**T.C.**

**ÇANAKKALE ONSEKİZ MART UNIVERSITY**

**SCHOOL OF GRADUATE STUDIES**

**DEPARTMENT OF PHYSICS**

# THE INVESTIGATION OF THE RELATIONSHIP BETWEEN THE PHYSICAL PARAMETERS OF SUNSPOTS AND SOLAR CORONAL MASS EJECTIONS USING ARTIFICIAL INTELLIGENCE ALGORITHMS

**MASTER OF SCIENCE THESIS**

**ABD-UR RAHEEM**

**Thesis Advisor**

**Prof. Dr. Hüseyin ÇAVUŞ**

**ÇANAKKALE – 2021**

T.C.
ÇANAKKALE ONSEKİZ MART UNIVERSITY
SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF PHYSICS

# THE INVESTIGATION OF THE RELATIONSHIP BETWEEN THE PHYSICAL PARAMETERS OF SUNSPOTS AND SOLAR CORONAL MASS EJECTIONS USING ARTIFICIAL INTELLIGENCE ALGORITHMS

MASTER OF SCIENCE THESIS

**ABD-UR RAHEEM**

**Thesis Advisor**
**Prof. Dr. Hüseyin ÇAVUŞ**

This study has been supported by TÜBİTAK.
Project No: 117F336

**ÇANAKKALE – 2021**

**T.C.**
**ÇANAKKALE ONSEKİZ MART ÜNİVERSİTESİ**
**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

The study entitled '**The Investigation of The Relationship Between The Physical Parameters of Sunspots and Solar Coronal Mass Ejections Using Artificial Intelligence Algorithms**', submitted by Abd-ur RAHEEM under the supervision of Prof. Dr. Hüseyin ÇAVUŞ was defended on 29/07/2021 has been approved unanimously as **thesis in Master of Science** in **Physics** of Çanakkale Onsekiz Mart University, School of Graduate Studies by the jury.

| **Members of Jury** | **Signature** |
|---|---|
| Prof. Dr. Hüseyin ÇAVUŞ **(Advisor)** | …… |
| Prof. Dr. Enise Nihal ERCAN **(Member)** | …… |
| Assist. Prof. Dr. Murat HÜDAVERDİ **(Member)** | ……. |

Thesis No : 10417427
Thesis Defence Date : 29/07/2021

……
Prof. Dr. Pelin KANTEN
Director of Enstitute

29/07/2021

## ETHICAL STATEMENT

I hereby declare that the work I have presented in this thesis, which I prepared in accordance with Çanakkale Onsekiz Mart University Graduate Education Institute Thesis Writing Rules, is original work and I have obtained the data, information and documents I have presented in the thesis within the framework of academic and ethical rules, that I have presented all information, documents, evaluations and results in accordance with scientific ethics and morals, that I have cited all the works I have used in the thesis by making appropriate references.

Abd-ur Raheem

29/07/2021

## PREFACE

First of all, I would like to express my gratitude to my advisor, Prof. Dr. Hüseyin Çavuş, who advised me in the realest sense of the word. Without his guidance and vision this work would not have been possible. Thank you for always providing academic freedom and encouraging me to think out of the box. I would like to thank Dr. Ahmet Cumhur Kinaci for answering all of my question related to machine learning and always sparing time for my inquiries, Prof. Dr. Haimin Wang and Prof. Dr. Jason T.L. Wang from NJIT for guiding me through the course of this study.

I cannot thank my father Dr. Azhar Waheed enough who has blessed me with his guidance, wisdom, and support on every turn of life and for being a role-model for me. I am forever in your debt, baba. I would like to express my gratitude to my loving mother Farah for her unconditional love and making sure that I ate enough for the day, my brother who has tried and succeeded to be there for my parents in my absence, and my sister who will surely be an amazing doctor.

I would like to thank my friend Gani Çağlar Çoban, with whom I share the office room, and enjoy endless interesting academic discussions. The idea behind this thesis originated in one of these thought-provoking discussions.

This study was supported by TÜBİTAK under project numbered 117F336. This study made use of the CDAW database which is provided by NASA through the CDAW datacentre in cooperation of the Naval Research Laboratory, data obtained from SOHO spacecraft which is a collaborative space mission between NASA and ESA, SHARP database which is hosted and curated by JSOC at Stanford University and version 2.1.2 of SunPy which is an open-source API for data access available in Python programming language.

Abd-ur Raheem

Çanakkale, July 2021

## ABSTRACT

## THE INVESTIGATION OF THE RELATIONSHIP BETWEEN THE PHYSICAL PARAMETERS OF SUNSPOTS AND SOLAR CORONAL MASS EJECTIONS USING ARTIFICIAL INTELLIGENCE ALGORITHMS

Abd-ur Raheem
Çanakkale Onsekiz Mart University
School of Graduate Studies
Master of Science Thesis in Physics
Advisor: Prof. Dr. Hüseyin ÇAVUŞ
15/06/2021, 56

Everyday magnetically active regions on the surface of the sun undergo various changes to cause Coronal mass ejections (CME) which in turn travel in the interplanetary environment sometimes causing shock waves within and are an important factor for space weather predictions and other related studies. The source active regions behind the initiation of CMEs until now are identified manually to our knowledge and thus the information regarding source active regions for CMEs is scarce and miniscule and source active regions regarding a vast majority of CMEs remain unknown. This study proposes a novel technique based on machine learning and other specific filtration processes to identify source active regions for a large number of CMEs and brings huge automation to the identification process. Magnetic data of previously known source Helio-seismic and Magnetic Imager (HMI) Active Region Patches (HARPs) for corresponding CMEs are used in study to train machine learning algorithms. The proposed mechanism includes the use of Long-Short term memory Networks (LSTM) specifically to simultaneously learn patterns in 17 magnetic parameters of HARPs collected from line-of-sight magnetograms and vector magnetograms of previously known source active regions. The produced LSTM model was then employed to identify source HARPs for all known CMEs during 2010 and 2020. Source active regions for 4895 CME events have been identified using this technique and a database of identified source active regions for CMEs is produced as a result of this study. Magnetic parameters of identified HARPs are then analysed with respect to corresponding CMEs to gain insight regarding the onset of CMEs.

**Keywords:** CME, Solar Active Regions, HARP, CME Onset, Source Active Regions, Space-Weather, Machine-Learning, Deep-Learning, Neural Networks.

## ÖZET

# GÜNEŞ LEKELERİNİN FİZİKSEL PARAMETRELERİ İLE GÜNEŞ KORONAL KİTLE ATEŞLERİ ARASINDAKİ İLİŞKİNİN YAPAY ZEKA ALGORİTMALARIYLA İNCELENMESİ

Abd-ur Raheem

Çanakkale Onsekiz Mart Üniversitesi

Lisansüstü Eğitim Enstitüsü

Fizik Yüksek Lians Tezi

Danışman: Prof. Dr. Hüseyin Çavuş

15/07/2021, 56

Güneş yüzeyinde bulunan manyetik olarak aktif bölgeler çeşitli değişikliklere uğrayarak, gezegenler arası ortamda şok dalgalarına, uzay hava durumu tahminleri ve diğer ilgili çalışmalar için önemli bir faktör olan Koronal Kütle Atımlarına (CME) neden olurlar. Şimdiye kadar CME'ler için kaynak aktif bölgeler bizim bilgimize göre manuel olarak tanımlandıkları için CME'ler için kaynak aktif bölgelere ilişkin bilgiler çok az sayıda ve CME'lerin büyük çoğunluğu için kaynak aktif bölgeler halen bilinmemektedir. Bu çalışma, çok sayıda CME için kaynak aktif bölgeleri belirlemek, makine öğrenimi ve diğer özel filtreleme süreçlerine dayalı yeni bir teknik önermekte ve tanımlama sürecine büyük bir otomasyon getirmektedir. Bu çalışmada ilgili CME'ler için önceden bilinen kaynak Helio-sismik ve Manyetik Görüntüleyici (HMI) Aktif Bölgelerin (HARP'ler) manyetik verileri, makine öğrenimi algoritmalarını eğitmek için kullanılmıştır. Daha önce bilinen kaynak aktif bölgelerin vektör ve görüş hattı manyetogramları kullanılarak HARP'lere ait 17 manyetik parametre verisi toplananmış ve bu HARP'lere ait bu manyetik parametrelerdeki değişiklikleri eş zamanlı olarak öğrenebilen bir Uzun-Kısa süreli bellek Ağı (LSTM) kullanılarak bir mekanizma önerilmiştir. Geliştirilen LSTM modeli daha sonra 2010 ve 2020 boyunca bilinen tüm CME'lere ait kaynak HARP'leri belirlemek için kullanılmıştır. Bu teknik kullanılarak 4895 CME olayı için kaynak aktif bölgeler belirlenmiş ve bunun bir sonucu olarak CME'ler için tanımlanmış kaynak aktif bölgelerin bir veri tabanı elde edilmiştir. CME'lerin ortaya çıkmasına ilişkin fikir edinmek için tanımlanan HARP'lerin manyetik parametreleri daha sonra ilgili CME'lere göre analiz edilmiş ve analiz sonuçları bu tezde verilmiştir.

**Anahtar Kelimeler:** Güneş Aktif Bölgeleri, Koronal Kütle Atımları, HARP, Kaynak Aktif Bölgeler, Uzay Hava Durumu, Makine Öğrenmesi, Derin Öğrenme, Sinir Ağları.

**TABLE OF CONTENTS**

## LIST OF ABBREVIATIONS

| | |
|---|---|
| CDAW | Coordinated Data Analysis Workshops |
| JSOC | Joint Science Operations Center |
| SHARP | Space weather HMI Active Region Patches |
| HARP | HMI Active Region Patches |
| AR | Active Region |
| CME | Coronal Mass Ejection |
| RF | Random Forest |
| DT | Decision Tree |
| SVM | Support Vector Machine |
| kNN | K-Neerest Neighbour |
| LSTM | Long-term Short-Term Memory |
| SOHO | Solar and Heliospheric Observatory |
| SDO | Solar Dynamics Observatory |
| L1 | Langrange 1 point |
| CCNN | Cascade correlation neural network |
| MLP | Multi-Layer Perceptron |
| HMI | Helioseismic and Magnetic Imager |
| MDI | Michelson Doppler Imager |
| SMART | Solar Monitor and Active Region Tracking |
| ANN | Artificial Neural Network |
| SF | Solar Flare |
| SEP | Solar Emission Particle |
| CAT-PUMA | CME Arrival Time Prediction Using Machine Learning Algorithms |
| ML | Machine Learning |
| AI | Artificial Intelligence |
| DL | Deep Learning |
| RNN | Recurrent Neural Network |
| PIL | Polarity-Inverted Lines |
| NASA | National Aeronautics and Space Administration |
| ESA | European Space Agency |
| CPA | Central Position Angle |

| | |
|---|---|
| KE | Kinetic Energy |
| LASCO C1 | Large Angle and Spectrometric Coronagraph (instrument 1) |
| LASCO C2 | Large Angle and Spectrometric Coronagraph (instrument 2) |
| LASCO C3 | Large Angle and Spectrometric Coronagraph (instrument 3) |
| MPA | Measurement Position Angle |
| LOS | Line Of Sight |
| LWS | Living With a Star |
| IGSO | İnclined Geosynchronous Circular Orbit |
| DONKI | Space-weather Database of Notifications, Knowledge, Information |
| NOAA | National Oceanic and Atmospheric Administration |
| CCMC | Community Coordinated Modelling Centre |
| HARPNUM | HMI Active Region Patch Number |
| API | Application Programming Interface |
| PA | Position Angle |
| ADAM | Adaptive Moment Estimation |
| TP | True Positive |
| FP | False Positive |
| TN | True Negative |
| TP | True Positive |
| F1 | F-score or F-measure |
| FOV | Field Of View |
| AAVSO | American Association of Variable Star Observers |
| SILSO | Sunspot Index and Long-term Solar Observations |
| MEANPOT | Mean photospheric excess magnetic energy density |
| TOTPOT | Total photospheric magnetic free energy density |
| USFLUX | Total unsigned flux |
| MEANGAM | Mean inclination angle |
| MEANGBT | Mean gradient of total field |
| MEANJZD | Mean vertical current density |

## LIST OF TABLES

identified in this study as a function of annual aggregation for each year.

## LIST OF FIGURES

# CHAPTER 1
# INTRODUCTION

Interest in understanding mechanisms behind the onset of solar events like solar flares, solar emission particles (SEPs), CMEs has increased in the last two decades mainly due to an increase in the availability of related data from space missions. This is partly due to the need to protect ever increasing man-made equipment in orbit and thus the increased importance and necessity of space weather prediction. The most important factor in this area is the understanding and comprehensive grasp over phenomena that arise from the continuously changing solar magnetic field. Active regions and their continuously changing complex magnetic field and its interaction with the interplanetary environment thus are one of the main and intriguing topics of research in solar physics and magnetohydrodynamics.

## 1.1 CMEs, Solar Flares, And Related Solar Phenomenon

CMEs are clouds of solar plasma along with electromagnetic radiation released from the solar photosphere in bursts of energy that propagate in interplanetary environment. CMEs generally have three distinct features namely, a core which is dense, a front edge which is relatively bright and a pocket which is low in electron density. The coronagraph images usually reveal this bright dense edge of the CMEs and is measurable and observable in magnetograms (See, Gopalswamy et al., 2009). CMEs are measured and observed through line of sight (LOS) magnetograms. CMEs when propagate through the interplanetary environment interact with the surroundings and the solar wind. CMEs may accelerate or decelerate after leaving the solar surface. Initially fast CMEs are usually decelerated with the reference of solar wind in its direction (Manoharan, 2006). Normally all the acceleration and deceleration of the released CMEs occur very close to the solar surface, but they are cases where this acceleration or deceleration is measured even past Earth's orbit (Freiherr von Forstner et al., 2018) and even around Mars (Richardson, 2014). CMEs within themselves may contain shock waves which occur when CMEs go past 500 km/s mark (Wilkinson, 2012). Parameters that include initial speed, acceleration, mass, and kinetic energy are calculated for these CMEs and are available in various catalogues including the one used in this study discussed in the methods section. CMEs are usually but not always accompanied with solar flares which are outburst of

electromagnetic energy usually at the sites of solar active regions. There is no exact one-to-one relationship between solar flares and CMEs although sometimes solar flares and CMEs take place together with each other. A lot of solar flares are associated with CMEs and many more are not (Webb and Howard, 2012).

### 1.2 Solar Active Regions And Their Relationship With Solar Phenomenon

Regions on the surface of the solar photosphere with relatively enrich magnetic field appear darker. This is due to the lower temperature of these regions due to relatively increased groupings of magnetic flux enriching the magnetic fields in these regions. This result due to convection. The temperature in these regions is roughly around 3000-4000K whereas the temperature at the solar photosphere is around 5780K therefore resulting in these regions appearing darker than the rest of the solar surface. Usually, these active regions that consist of structures known as sunspots or collectively sunspot groups, have inversed magnetic polarity within themselves. These phenomena are also observable on stars other than the Sun and are called starspots (Strassmeier, 1999). The total number of active regions on the solar surface at any given time is dependent on the state of the solar cycle. A solar cycle normally spans around 9-11 years during which there is a period of maximum activity i.e., an increased number of active regions come to existence and disappear. Active regions constantly change in structure as the complex magnetic structures of these active regions' changes, also they undergo magnetic reconnections and disintegrations to appear or disappear from the solar surface. Normally it takes around 10-11 days for an active region to travel through the observable solar surface from one end to another depending on its latitude on the solar surface as the solar surface rotates at different speeds at different latitudes, a phenomenon known as differential rotation (Weiss, 1965). This is observable through the solar active regions databases also discussed in the method section of this thesis. LOS and vector magnetograms both are used to detect and record the electromagnetic and magnetic field data of active regions. LOS magnetograms only contain visual features whereas vector magnetograms contain the parameters related to the magnetic field of active regions.

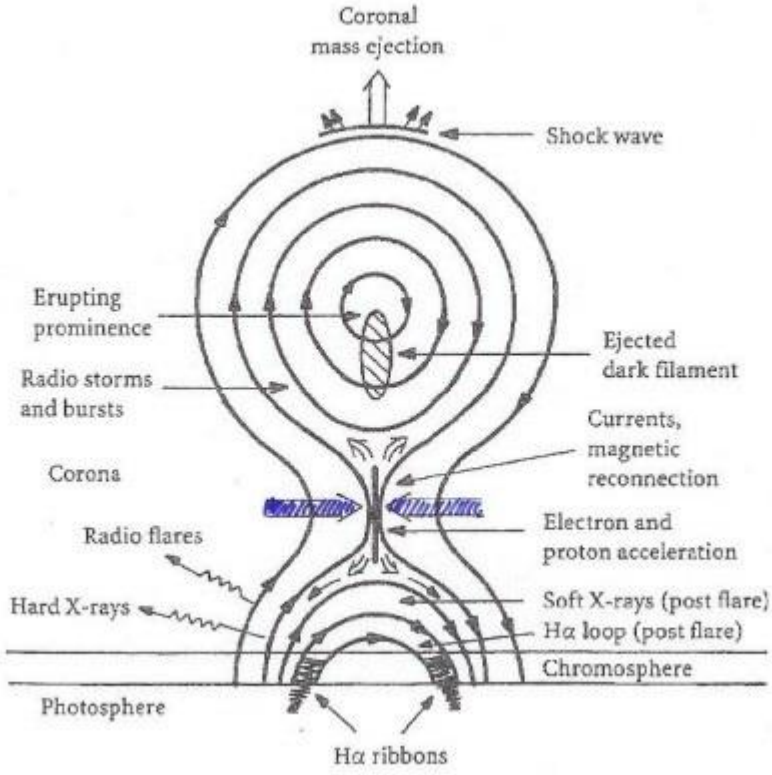



Figure 1. Schematic diagram of an CME eruption along with a solar flare caused by the magnetic reconnection occurring in the solar surface (Lang, 1995)

Active regions on the solar surface are the main sources behind many of the solar phenomenon including but limited to solar emission particles, solar flares, and CMEs. The complex magnetic structures of these regions cause magnetic loops to form around and within these active regions. These magnetic loops are usually very complex magnetically, visible and their heights depends on their respective magnetic complexity. These magnetic loops can undergo processes such as magnetic reconnection or disintegrate to cause CMEs, solar flares, and other solar phenomena (Figure 1). Magnetic reconnection can be categorized as a phenomenon where magnetic field lines that are opposite instantaneously rearrange or re-order which swirls solar plasma and electromagnetic radiation into the interplanetary environment in the process. The magnetic field lines become intertwined and twisted during magnetic reconnection processes and eventually break to release CMEs and electromagnetic radiation (Green, 2016).

### 1.3 Why Is The Identification Of Source Active Regions A Necessity?

Many studies have found increasingly positive correlation between CMEs and the magnetic field parameters of source active regions, and it is found that parameterization of magnetic field of active regions help forecast and study phenomenon including but not limited to solar flares and CMEs (Schrijver, 2007; Mason and Hoeksema, 2010). Magnetic field parameters which represent measures of the non-potential magnetic field of the active regions are particularly important in this regard and are directly correlated with the solar activity specially the onset of solar flares and CMEs (Falconer et al., 2003; Leka and Barnes, 2003(a)). Any research regarding solar phenomenon including but not limited to CMEs and solar flares fundamentally requires data related to the changes in the magnetic field of the source active regions since nearly all solar phenomena originate from changes in the magnetic structures of active regions and their interaction with each other and surrounding interplanetary environment. So, it is necessary to model and understand the complex magnetic field and its variation with respect to time and processes originating from its interactions comprehensively. Solar phenomenon like CMEs and interplanetary shocks are extremely important factors in the space weather studies and play an important role in fields dealing with satellite operations, space instrument deployment and maintenance of space instrument as well as weather prediction operations on Earth.

### 1.4 Databases And Space Missions That Made This Thesis Possible

The Solar and Heliospheric Observatory (SOHO) was launched in 1995 as a collaboration between European Space Agency (ESA) and the National Aeronautics and Space Administration (NASA). The Large Angle and Spectrometric Coronagraph (LASCO) and Michelson Doppler Imager (MDI) onboard the spacecraft were essential for this study as the CME database, Coordinated Data Analysis Workshops (CDAW) uses these for the identification and calculation of the CME parameters. CDAW lists all the CME events from 1996 till date along with CME parameters such as positional angles, initial linear speed, the second order speed of CMEs at various heights, mass, acceleration, and the kinetic energy of the CMEs. NASA's Solar Dynamics Observatory (SDO) was launched in 2010 and houses the Helioseismic and Magnetic Imager (HMI) that is used by Stanford university's Joint Science

Operations Center (JSOC) to track and identify HARPs on the solar surface and their database Space weather HMI Active Region Patches (SHARP) (Bobra et al., 2014) contains the magnetic data for these HARPs which was used in this study to fetch data for the active regions associated with CMEs to make datasets for training ML algorithms used in the scope of this thesis. The space weather Database of Notifications, Knowledge, Information (DONKI) was crucial this study as the associated cases of the CMEs and active regions were taken from there upon which the machine learning algorithms were trained and later used for the identification of the source active regions of respective CMEs. The database lists known cases of the associations of active regions and CMEs, solar flares and other solar phenomena which is reported by the literature in this field. These associations are usually performed manually to our knowledge.

### 1.5 Machine Learning Techniques And The Motivation Behind This Thesis

Machine learning algorithms represent techniques that mimic human brain in its capability to understand and learn patterns and other useful relationship in data for forecasting and predictive applications. These algorithms essentially learn relationships in provided data and predict the outcome or output depending on the use case. Machine learning algorithms including Decision Trees, Support Vector Machine (SVM), Random Forest (RF), k-nearest neighbour (kNN) and Long-Short Term Memory networks (LSTM) are used in this study to identify patterns parameters of magnetic field of active regions with respect to the CMEs initiated by these regions. Decision tree algorithms are methods that display decision processes in a tree-like model of possibilities based on change in input data and rate these possibilities, their consequences, and the relationship of these consequences with each other to interpret an event or process and predict its outcome (Kumar and Sharma, 2016). RF classifiers, usually used for classification tasks, are ensembles of large number of decision tree models that work together co-dependently to model an event or process and predict possible outcomes of that event or process (Shi and Horvath, 2006). SVMs are algorithms which learn difference between different classes present within data by finding a hyperplane in N-number of dimensions that divides the data into different planes (Sanz, 2018). Artificial neural networks consist of network of data processing units called neurons which are tied together with each other in a network to model an event based on quantifying data related to that event and learn the relationship of

these data in correlation to the outcome by changing weights and biases in these neurons by repetitive predictions of the output data. LSTM networks are a form of recurrent neural networks which consist of cells that are assigned memory that hold information for an arbitrary interval of time related to previous predictions and gates (input gate, output gate and forget gate) are used to tune the amount of information or data flowing in or out of these cells (Graves and Schmidhuber, 2005). LSTM networks have been proven to be exceptionally excellent at learning patterns in large time-series data and are advantageous in cases where insensitivity to time gaps in necessary (Karim et al. 2019).

Since the data from vector magnetograms that include magnetic field parameters of active regions form a set of complex large time-series machine learning and deep learning is used in part in the scope of this thesis to train ML models the patterns behind the initiation of CMEs with respect to changes in these co-dependant magnetic parameters. To our knowledge, no large-scale method for identification of source ARs with regards to CMEs is present and source ARs for a very small number of CMEs have been identified. Currently the identification of source active regions is performed manually as per our knowledge. The motivation of this study is, therefore, to purpose a mechanism for the identification of source ARs for a large numbers of CMEs automatedly, with help of various machine learning algorithms and other filtration procedures and study the properties of source ARs with respect to the CMEs they initate. A mechanism has been proposed and applied as a result of this study and some of the relationships between the magnetic field parameters of source ARs and the CMEs initiated by them have been studied and results are presented in this thesis.

The next chapter lists the previous studies in literature in this field. Chapter 3 named as Material and Method explains in detail the data products used in this study and their properties along with the steps and processes performed on these data products to achieve the goals of this study. Material and Method also elaborates the method including the filtration procedure and the design and architectures of the machine learning algorithms used. In the 4th chapter the results are given as graphs and tables. Finally, the last chapter goes in details about the results and discussion on the results and then the conclusions derived from the analysis of the properties of identified source ARs with the CMEs initiated by them.

# CHAPTER 2
# PREVIOUS STUDIES

## 2.1 Interest In Solar Activity Prediction And Pioneer Studies

The investigation of solar activity especially solar flares and CMEs has peaked due to an interest in space weather which can be attributed to an increased presence of the man-made equipment and spacecrafts in orbit thus producing both data to study such events and also the need to protect these instruments by understanding the onset of such solar events and monitoring the space weather by collecting and manipulating data for this goal. The most important phenomena in the solar activity in terms of its impact on Earth and its orbits are CMEs followed by solar flares (Gosling, 1993; Hudson, 1995). It can be said that the likelihood of the initiation of CMEs increases as the magnetic complexity of the active regions increases (Michalek and Yashiro, 2013) as CMEs are more likely to occur from more mature i.e., magnetically more complex active regions. This is why the relationship of the magnetic field parameters of the active regions and solar activity is very obscure and difficult to interpret. Shock waves that occur in the CMEs are primarily responsible for the disturbances in the orbital instruments and sometimes the weather on earth. Cavus et al. (2020) studied the correlation between the sunspot number and the frequency and intensity of shock waves from 1995 till 2018 and found good correlations between these two solar phenomena.

## 2.2 Parameterization Of The Magnetic Field Of Solar Active Regions

Studies quantifying the non-potentiality of solar active regions has produced some interesting and important results over the years. Especially the spatially averaged signed shear angle parameter has been discussed in many studies and reported to be an important quantifying parameter for non-potentiality of active regions and thus highly correlated with the solar activity including solar flares and CMEs (Falconer et al., 2008; Tiwari et al., 2010; Wang et al., 2002). Falconer et al. (2008) studied the measures of vector magnetograms for non-potentiality in reference to the solar activity predictions. They concluded that parameters including the length of strong-shear main neutral line and the length of strong-gradient main

neutral line correlate positively with solar activity. It is well established till this point that multipolar active regions produce more solar activity i.e. solar flares and CMEs than bipolar active regions (Zirin, 1988) and as these mentioned parameters correspond to measurement of the complexity of the active regions, the study produces good correlation with solar activity. Tiwari et al. (2010) studied four different active regions and over a hundred samples of vector magnetograms and concluded that the lower limit of the spatially averaged signed shear angle parameter to be an important and decisive feature for solar flare prediction. Their analyses show a dependence of the intensity of solar flares on the magnitude of the spatially averaged signed shear angle parameter of that particular active region. Nindas et al. (2003) studied six active regions that were associated with geomagnetic storms and the initiation of CMEs and concluded that the magnetic helicity by transient photospheric flows in these regions correlate positively with some of the CME parameters but cannot be separately used to predict the initiation of CMEs.

Studies have already demonstrated that the parameterization of the magnetic field parameters of the active regions positively correlates with the solar activity and thus with the prediction of the activities on solar surface (Falconer et al., 2003; Leka and Barnes, 2003(b); Schrijver, 2007). The parameterization of the magnetic field of active regions from HMI instrument onboard the SDO spacecraft started automation of the tracking, collection of data and extraction of these quantities. Automated systems for tracking solar active regions and their magnetic properties have been proposed and implemented for major space missions in this field. Higgins et al. (2011) proposed and implemented an automated system called Solar Monitor and Active Region Tracking (SMART) for the detection, tracking, and cataloguing of solar active regions using the MDI instrument onboard SOHO. The performance of the automation process and its details on the processes involved can be obtained from Hoeksema et al. (2014) which is authored by the team behind SHARP database. This database helped pave the way for studies related to the prediction and investigation of the onset of various solar activity including solar flares and CMEs.

## 2.3 Usage Of Machine Learning To Predict Solar Activity

Solar flares can be reliably predicted beforehand within different time intervals (Inceoglu et al., 2018) with varying level of confidence. Studies regarding solar flare prediction and forecasting that use the X-ray timings and various machine learning algorithms are widespread in solar physics and space weather studies. Various studies have made use of machine learning techniques including but not limited to RF, SVM, ANNs and LSTMs for the prediction of solar flares along with studies regarding the correlations between the change in magnetic field parameters of the source ARs and solar activity especially solar flares (Qahwaji and Colak, 2008; Bobra and Ilonidis, 2016; Liu et al., 2018; Liu et al., 2019). Qahwaji et al. (2008) used machine learning algorithms to check for the relationship between some CMEs and their associated flares. They used SVM and Cascade correlation neural network (CCNN) in particular to predict whether a flare will initiate a CME using the flare properties of the flare as input. The two machine learning algorithms are compared, and recommendations are made in favour of using a more advanced machine learning technique. Their use of jack-knife technique to determine the bias in the magnetic field parameters of the solar phenomena encourages use of more advance machine learning techniques to catch patterns in data of the magnetic field parameters which is a time-series data. Bobra and Ilonidis (2016) pioneered the usage of ML algorithms with regards to magnetic field data provided by SHARP to predict for CMEs, solar flares, and other solar phenomena. They used SVMs to predict weather a flare will also initiate a CME or not by feeding the model with populations of data that are classified beforehand to learn for distinguishing values of the used parameters. Inceoglu et al. (2018) used two machine learning techniques namely, SVM and multilayer perceptron (MLP) to distinguish between flares with associated CMEs and SEP from flares and CMEs which are not associated with any other event. They compare the two algorithms and show that SVMs are slightly better at predicting these phenomena. They also point out the need of longer time-series data as is by this study because the amount of data greatly effects the ability of machine learning algorithms to generalize data. Yang et al. (2018) attempted predicting the speed of the solar wind up to 1 AU using ANNs demonstrating in the process that ANNs are a viable solution to the prediction problems related to solar activity. Their models were able to predict the fluctuations of solar winds at distances up to geostationary orbits of Earth.

## 2.4 State Of Machine Learning In The Prediction Of Solar Activity And The Novelty Of Study

Liu et al. (2018) demonstrated the use of their CME arrival Time Prediction using machine learning Algorithms (CAT-PUMA) to predict for the arrival times of halo and non-halo CMEs. The technique uses a dataset of previously labelled CMEs along with their parameters including CME speeds along with a set of machine learning algorithms based on SVM technique. The new approach manages to predict the CME arrival times within a net error of 6 hours. The motivation being investigation of CMEs in the interest of space weather prediction. Liu et al. (2019) recently presented a study that used LSTM networks to predict whether a flare producing active region will also produce a CME. Their technique uses a parameter ranking mechanism to rank the magnetic parameters of the active regions and then using the magnetic parameters with regards to their effectiveness scale on RNNs and LSTM networks to predict weather a flare producing active region will also produce a CME using a small list of previously known source active regions listed in the DONI database. Liu et al. (2020) demonstrated the use of RNN (a machine learning technique that makes use of time-series data to learning patterns with respect to their change in time) to distinguish solar flares that initiate CMEs from those which do not. The model works with M-class and X-class flares and predicts its potential with respect to the initiation of CMEs. There is room for improvement in the quality of the data available at SHARP for the magnetic parameters of solar active regions as Wang et al. (2019) demonstrated an increase in the predictiveness of the magnetic parameters using machine learning techniques. They used methods including weighing the magnetic parameters with their polarity-inverted lines (PIL) and recorded an increase in the performance of machine learning algorithms in terms of predictiveness and a decrease in false positives of these models.

This study in contrast to the literature in this field proposes a new mechanism that uses novel filtration processes along with machine learning techniques for the identification of source ARs for all the CMEs that have been recorded in the CDAW database for CMEs. To our knowledge no attempt has been made to automate the identification process of the active regions for CMEs before. The database which is produced as a result of this thesis provides a unique opportunity to study the relationship of the change in magnetic field of active regions

with respect to initiation of CMEs. An analysis has also been done regarding this relationship and results are presented in the results section.

The next section in this thesis goes over the materials used in this study and the method proposed and implemented through the course of this study.

# CHAPTER 3
# MATERIAL AND METHOD

## 3.1 Material

### 3.1.1 Data Regarding Cmes And Their Properties

The data of the CMEs used in the study during the training procedure of ML algorithms and then for identification processes were taken from the CDAW database hosted by NASA(URL-1). CDAW collects and hosts CME data obtained from the Large Angle and Spectrometric Coronagraph (LASCO) instrument on NASA's Solar and Heliospheric Observatory (SOHO) spacecraft. SOHO is an international collaboration between NASA and ESA. The instrumentation onboard is mutually and collaboratively designed and produced both in USA and Europe. SOHO is an orbit at the Lagrange L1 point between Earth and the Sun approximately 1.5 km from the Sun. There are, at the time of writing this thesis, around 14604 CMEs listed in the CDAW data centre. The data regarding each CME is given below in the table as an example and properties including CME linear speed, acceleration, mass, and kinetic energy are given in CDAW database.

Table 1

An example of the CME parameters obtained from CDAW. Here CPA is the abbreviation for central position angle whereas MPA denotes the measurement position angle.

| **Timestamp** | **CPA** | **Width** | **Linear Speed** | **2nd order initial speed** | **2nd order final speed** | **2nd order speed at 20R** | **Accel** | **Mass** | **K.E.** | **MPA** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | deg | $kms^{-1}$ | $kms^{-1}$ | $kms^{-1}$ | $kms^{-1}$ | $ms^{-2}$ | g | erg | deg |
| 2007/05/31 13:30:04 | 234 | 17 | 238 | 204 | 276 | 290 | 2.0 | - | - | 240 |
| 2007/05/31 18:31:41 | 104 | 15 | 253 | 159 | 362 | 387 | 5.4 | - | - | 105 |
| 2007/06/01 00:06:04 | 227 | 13 | 287 | 345 | 224 | 70 | -5.5 | - | - | 236 |
| 2007/06/01 05:06:05 | 248 | 14 | 311 | 252 | 370 | 716 | 18.7 | - | - | 252 |
| 2007/06/01 07:30:04 | 88 | 53 | 337 | 415 | 250 | 0 | -8.8 | 1.8e+14 | 1.1e+29 | 93 |
| 2007/06/01 14:06:04 | 228 | 7 | 161 | 142 | 178 | 315 | 3.6 | - | - | 235 |

Normally CPA and MPA are same, but they can be slightly different for HALO CME in particular because of the difficulty of calculating position angles for them due to perspective of view of the spacecraft being at L1 (Alex et al., 2006). The energy for the CME is calculated based upon parameters including linear speed and mass of the CME (Gopalswamy et al., 2009). The calculations for mass and kinetic energy for some instances are not available in the database and thus these events were ignored in this analysis performed after the learning procedure finished and identifications were done. Comments are added for each event in the CDAW database reflecting the quality of the CME measurements. CDAW measures, combines, and averages CMEs from the three coronagraphs present on the LASCO instrument namely, LASCO C1, LASCO C2 and LASCO C3.

### 3.1.2 Magnetic Field Data For Active Regions

The magnetic field data of solar active regions was required in this study since, as mentioned in the previous section, changes in the magnetic fields of solar active regions trigger solar events including but not limited to CMEs and SFs. This magnetic data comes from vector magnetograms which then have to transformed into usable coordinate systems and filtration processes like noise suppression, coordinate compensation, calibration correction has to be applied. Stanford university's JSOC data centre is useful and praise-deserving in this regard. JSOC consolidates vector magnetogram data from different spacecrafts and applies mentioned calculations for the data to be ready to use for research purposes. This study used the data products derived from the Helioseismic and Magnetic Imager (HMI) onboard NASA's SDO, first spacecraft in the long-term Living with a star (LWS) project. SDO launched in 2010 and is currently placed in an inclined geosynchronous circular orbit (IGSO) which is approximately 35,756 km above Earth's surface.

### 3.1.3 HMI Active Region Patches And JSOC

JSOC provides data products from other space missions along with HMI. The SHARP database for the HMI data products encapsulates a program that identifies, tracks, and stores magnetic parameters of the non-potentiality of active regions. There are two main time-series

for magnetic field parameters of non-potentiality of solar active regions in the SHARP database namely, hmi.sharp_720s and hmi.sharp_720s_cea and their near time versions namely, hmi.sharp_720s_nrt and hmi.sharp_cea_720s. The definitive time-series have more certainty and the near real time time-series have more recent data but low certainty. SHARP denotes the active regions as patches and names them HMI Active Region Patches (HARPs). The change comes as the definitive changes are processed as some HARPs merged together with each other and form a greater HARP and sometime HARPs disintegrate and form smaller HARPs. So, the definitive time-series includes data that is not certain, but all the calculations have been performed and thus the name definitive. The near real time time-series on the other hand is contains more recent data but the data is subjected to change. The hmi.sharp_cea_720 data contains magnetic data that has been remapped and projected to a cylindrical equal area in the cartesian coordinate system. The hmi.sharp_cea_720s time-series is used in this study. This data is also cantered on AR that is being tracked. The quantities provided in the SHARP database include the velocity, continuum intensity, line of sight magnetic field including 31 other magnetic field quantities that include the magnetic flux density, azimuth angles and coordinates of the HARPs among other. An overview of the magnetic parameters used in this study is given in Table 2.

Table 2

An overview of the magnetic field parameters of HARPs used in this study during the training of machine learning algorithms and then for identification.

| **Parameter** | **Description (unit)** | **Formula** |
|---|---|---|
| MEANPOT | Mean photospheric excess magnetic energy density (erg/cm$^3$) | $\overline{\rho} \propto \frac{1}{N}\Sigma(B^{Obs} - B^{Pot})^2$ |
| TOTPOT | Total photospheric magnetic free energy density (erg/cm) | $\rho_{tot} \propto \Sigma(B^{Obs} - B^{Pot})^2 dA$ |
| USFLUX | Total unsigned flux (Mx) | $\Phi = \Sigma\lvert B_z\rvert dA$ |
| MEANGAM | Mean inclination angle (Deg) | $\overline{\gamma} \propto \frac{1}{N}\Sigma \arctan(\frac{B_h}{B_z})$ |
| MEANGBT | Mean gradient of total field (G/Mm) | $\lvert\overline{\nabla B_{tot}}\rvert = \frac{1}{N}\Sigma\sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}$ |
| MEANJZD | Mean vertical current density (mA/m$^2$) | $\overline{J_z} \propto \frac{1}{N}\Sigma(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y})$ |
| TOTUSJZ | Total unsigned vertical current (A) | $J_{Ztotal} = \Sigma\lvert J_z\rvert dA$ |
| MEANGBH | Mean gradient of horizontal field (G/Mm) | $\lvert\overline{\nabla B_h}\rvert = \frac{1}{N}\Sigma\sqrt{\left(\frac{\partial B_h}{\partial x}\right)^2 + \left(\frac{\partial B_h}{\partial y}\right)^2}$ |
| MEANGBZ | Mean value of the vertical field gradient, in G/Mm | $\lvert\overline{\nabla B_z}\rvert = \frac{1}{N}\Sigma\sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}$ |

| | | |
|---|---|---|
| MEANALP | Mean twist parameter, $\alpha$ ($Mm^{-1}$) | $\alpha_{total} \propto \frac{\Sigma J_z B_z}{\Sigma B_z^2}$ |
| MEANJZH | Mean current helicity ($G^2$/m) | $\overline{H_c} \propto \frac{1}{N}\Sigma B_z J_z$ |
| TOTUSJH | Total unsigned current helicity ($G^2$/m) | $H_{Ctotal} \propto \Sigma \lvert B_z \cdot J_z \rvert$ |
| ABSNJZH | Absolute value of the net current helicity ($G^2$/m) | $H_{C\ abs} \propto \lvert \Sigma B_z \cdot J_z \rvert$ |
| SAVNCPP | Sum of the absolute value of the net current per polarity (A) | $J_{Zsum} \propto \lvert \Sigma^{B_z{}^+} J_z dA \rvert + \lvert \Sigma^{B_z{}^-} J_z dA \rvert$ |
| MEANSHR | Mean shear angle (Deg) | $\overline{\Gamma} = \frac{1}{N}\Sigma \arccos\left(\frac{B^{\mathrm{Obs}} \cdot B^{\mathrm{Pot}}}{\lvert B^{\mathrm{Obs}} \rvert \lvert B^{\mathrm{pot}} \rvert}\right)$ |
| SHRGT45 | Area with shear angle greater than 45 degrees (percent of total area) | Area with shear greater than $45^{\circ}$/total area |
| R_VALUE | Flux along gradient-weighted neutral-line length (Mx) | $\Phi = \Sigma \lvert B_{LOS} \rvert dA$ within $R$ mask |

### 3.1.4 DONKI And Previously Identified Source Active Regions

The DONKI database (URL-2) also hosted by NASA at The Community Coordinated Modelling Centre (CCMC) lists CMEs and Solar Flares (SF) with identified source active regions. The Moon to Mars Space Weather Operations Office provides data and the and indexes solar events like CME, SF, SEP etc among other which is then consolidated by DONKI (Berkebile-Stoiser et al., 2012). This provides a unique opportunity in regard to training machine learning algorithms which can be then used for identification of the source ARs later on hence the idea behind this thesis. There are 156 CMEs events in total with identified source ARs identified with a NOAA number. Of these events only 120 events are usable as the magnetic data for rest of the active regions is not available in SHARP database. The source active regions available in the DONKI database has NOAA number identifier for every listed event. In SHARP database for the HARPs NOAA identifier for that specific active region patch is available where applicable but since SHARP identifies and tracks more active regions than most other databases, NOAA number identifier for some of active region patches are not available. There is an HARPNUM identifier for each HARP identified and tracked across the solar surface. This occurs partly due to labelling procedure of SHARP database as SHARP labels HARPs after the patches have appeared and the either disappear or crossed the solar surface. This is done in this manner because some active region parches over the course of their lifetime merge with other active region patches and form larger active regions. On the other hand, in some active region patches may disintegrate into smaller active region patches. So, to combat these cases SHARP take these considerations into account before labelling the active region patches. Whereas NOAA creates a new NOAA number for each new identified active

region. So, in some cases there could have been conflicts in the NOAA-given NOAA identifier and HARP-given NOAA identifier of some active regions. HARPNUM identifier avoids these conflicts. SHARP puts NOAA identifiers on the active regions identified by NOAA but gives all self-identified active regions a HARPNUM.

By deriving the NOAA number identifier for the source active regions corresponding HARPNUM of these active region patches were identified and magnetic field parameters of these active regions were obtained from SHARP. The data from SHARP was attained using the API provided by JSOC (also available through the python library known as SunPy). SunPy provides a framework to fetch and query data from datacentres including but limited to JSOC, HINODE etc. (Mumford et al., 2015).

## 3.2 METHOD

### 3.2.1 Steps Of The Mechanism Developed And Deployed In This Study

The mechanism can be divided into the steps given in the following and a flowchart of the whole procedure as deployed in this study is given in figure 2.

I Firstly, data for the HARPs are collected and datasets are formed. This step includes calculation of position angle (PA) and labelling of the datasets using a technique by which a pulse parameter is created for each time series within a dataset discussed in section 3.2.2. The datasets are normalized and then padding is adjusted for each parameter in the dataset separately. The datasets are then transformed into a time-series format used in machine learning and described in section 3.2.3 till 3.2.5.

II Machine learning models are trained on the labelled datasets obtained with the help of DONKI, SHARP and CDAW. In this step models are trained using Decision Tree, SVM, RF, kNN and LSTM on all the datasets of different time span separately to determine which algorithm is best suited for the problem at hand. LSTM proves to be the best performing machine learning technique for this task. The details of this step are given in section 3.2.6 till 3.2.9. After the model is well trained and performs in an acceptable manner, it is deployed to identify source AR for all previously known CMEs between 2010 and 2020 using AR-Pools of these CMEs. AR-Pools are discussed in the next step.

III Before the deployment of the LSTM model for identification a custom filtration process is employed to form AR-Pool for every CME which consists of all potential-source ARs for that particular CME. Since, we have CME data from CDAW between 1996 and 2020. We also possess magnetic field data from SHARP of all identified active region patches from 2011 till 2020. This helps to decrease time-consumption for the whole process and also reduces the active region candidates for a CME that can never be source for that CME. This step greatly boosts the number of correct predictions made by the model later on. This step is discussed in detail in section 3.2.10.

IV The LSTM model is given the magnetic field data of the ARs in the AR-Pool of the CME and a prediction for CME is made by the LSTM model. If this prediction matches with CME present in CDAW with respect to time then, the predicted AR is labelled source AR for that CME. In this manner the model actually chooses an AR from the AR-Pool of the CME in question. This step is discussed in detail in section 3.2.10. Any predictions where more than one AR is picked by the model from the AR-Pool is considered incorrect prediction and discarded.

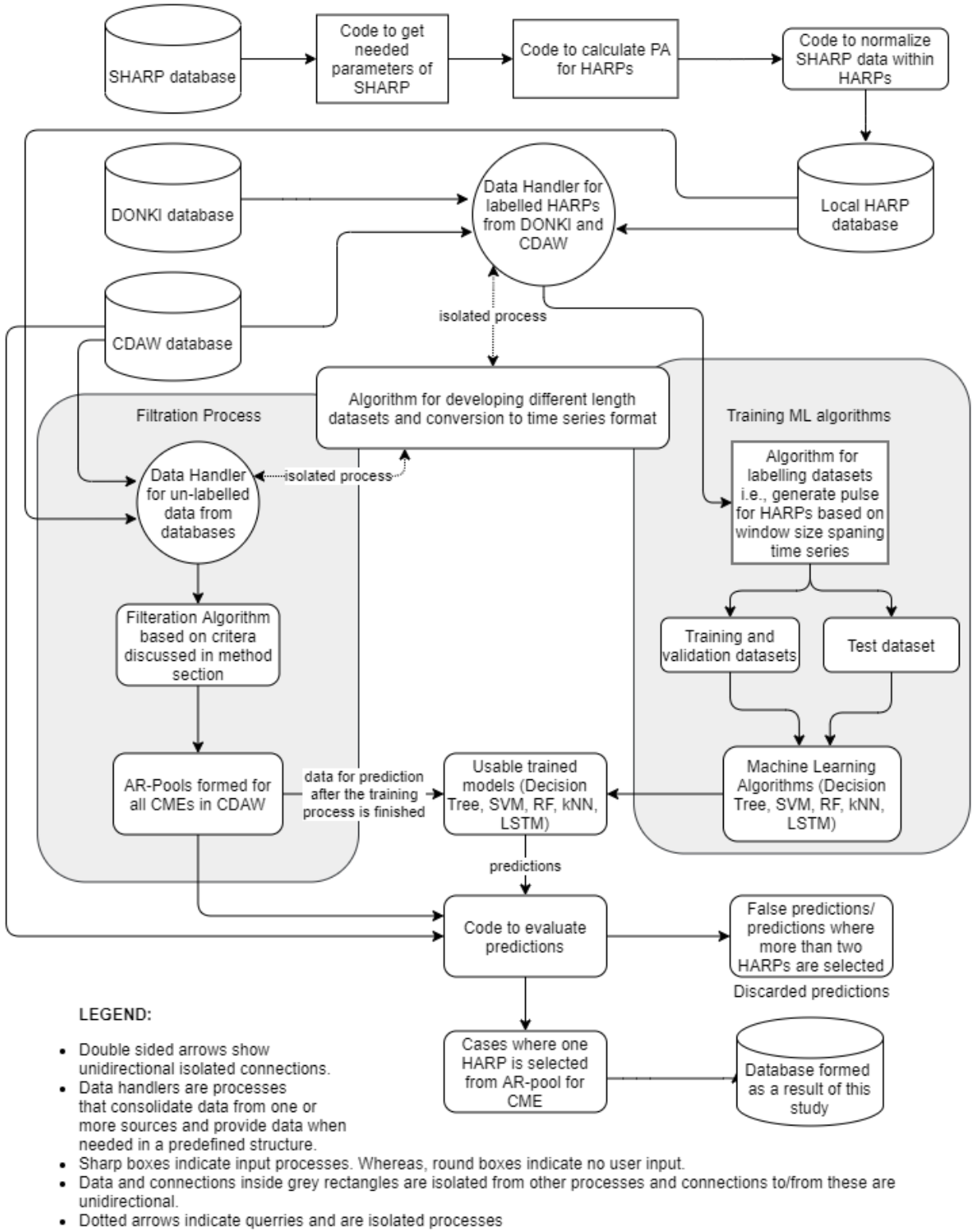



Figure 2. A flowchart detailing the whole procedure as developed and deployed in this study. The legend describes in the figure describes the different arrows used and different box types used in the flowchart.

### 3.2.2 Preparation Of Datasets And Their Normalization

Six different datasets are designed in the course of this study for each HARP. There are the 2h-dataset, 4h-dataset, 6h-dataset, 8h-dataset, 10h-dataset, and 12h-dataset. The only difference is the time span for the data of the 17 magnetic field for each HARP. These are designed so that the corresponding CME usually falls in the centre. The magnetic field parameters are obtained from SHARP through SunPy along with the maximum and minimum latitude and longitude at each time step. The cadence of SDO is approximately 12 minutes. So, the time step at a 12-minute interval. The position angle (PA) is calculated for each timestep of the dataset. This is performed by a custom code written to convert the latitude and longitude of the HARP at the solar surface to PA. After this step the datasets are normalized within each HARP between -1 and +1. This is an important step and by normalizing the data within each HARP has two reasons. Firstly, the difference in the magnetic parameters before and after the CME is of utmost importance and this is what the machine learning model is expected to learn. This helps the models to learn the changes in the magnetic field parameters rather than memorizing their particular values at any instance. Secondly, there are time-gaps in the datasets naturally since the identified source regions for CMEs are apart from each other in time. There are only 120 usable events in the span of ten years. Machine learning algorithms specially LSTM and other recursive neural networks are very sensitive to time gaps. Normalizing the data in this manner helps avoids this time gaps problem. Since, now when the data of a HARP ends and new one starts the values do not jump radically. This also help with the exploding gradient problem where the models are unable to learn from a time-series data if the values jump radically in between a time-series.

### 3.2.3 Padding The Datasets And Filling Time Gaps In The Datasets

The datasets are then adjusted for some absent data in some HARPs. This is done by padding the data at the end of each HARP where applicable. In some cases, there is missing data in between present data. So, the data jumps from the 06:12:00 to 06:36:00 instead of the normal 12-minute interval of 06:24:00. In these cases, average of data after and before this missing step is calculated and the filled in between the data. The amount of such filler data is along with the padding added in some cases is less than a total of 3%. This step is necessary as

the used machine learning models are very sensitive to time-gaps and the missing time steps and to create exceptions for each of these cases is very time-consuming and produces less accurate results in the long run.

### 3.2.4 Labelling Datasets By Creating Pulse Parameter For Each Dataset

The labelling procedure of each dataset is performed by creating a custom parameter called 'pulse' for each HARP in the different datasets. The pulse is a binary parameter which reflects the presence of CME in the time-series of the HARP at that moment. It takes value of 1 if there is a CME at that moment and 0 when there is no CME in the CDAW database. 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. So, the number of datapoints increase in the datasets (from 2 hours to 12 hours). This way pulse parameter itself becomes a time-series which holds the information about the presence and absence of CME and peaks to show the presence of a CME within the dataset. The 17 magnetic parameters make 17 different but co-dependant time-series within each dataset for each HARP and pulse is the $18^{th}$ parameter that is the target of the machine learning algorithms. By adding the pulse parameter now this becomes a classification problem where pulse value 1 depicts the CME class and pulse value represent NOTCME class. This binary classification has to be performed at each timestep by considering the changes in the 17 different time-series within the HARP at any time. This is very complex problem because all the 17 time-series are co-dependant on each other, and their mutual relationship causes the onset of CME. This problem is not only complex for human but also for machine learning algorithms as the number of variables and the possibilities becomes overwhelming.

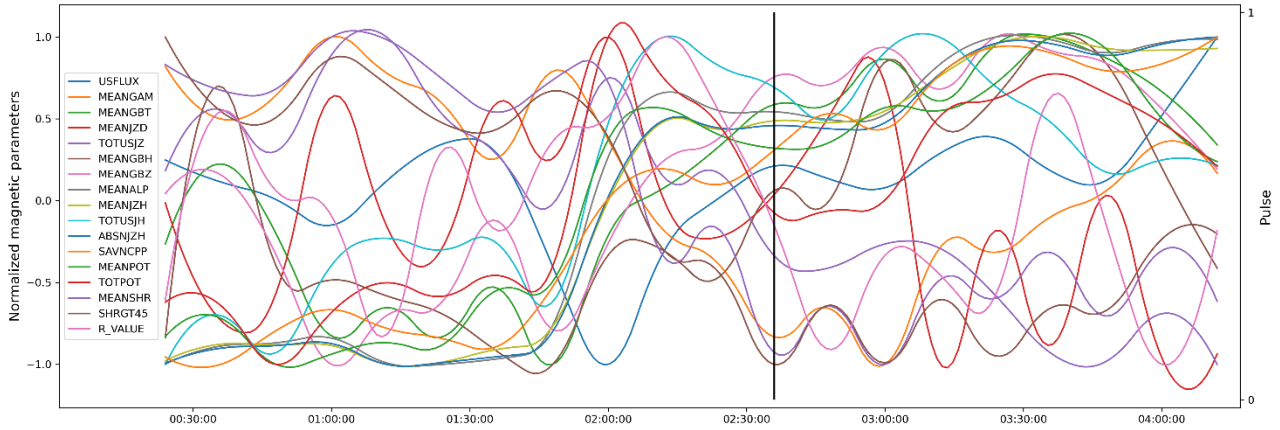



Figure 3. A representation of the first HARP of the 4-hour dataset along with the pulse parameter generated for it based on CME timing from CDAW database.

During the process of creating pulse for the HARPs in the datasets there should be no CME where the pulse takes the value of 0. As this is a time interval, in some cases there are other CMEs in the CDAW database during the time-period shown in the graph at bottom in the figure within the same positional angle constraints ($\pm 90^0$ of mean positional angle) as the CME used corresponding to the HARP used. So, in these cases the value of '0' and in turn the label 'NOTCME' for that time stamp cannot be applied with certainty. Such HARPs were discarded. This was later found to be very useful step as the performance of the models significantly increased after such samples were discarded for the datasets. One caveat is that the amount of data for larger time-interval datasets like (10-hour, 12-hour) consequently reduced because more such cases (mentioned before) emerge as the time interval increases. Table 3 shows the number of the HARPs and CMEs in each dataset. As observed from the third column the number of used HARPs decreases as the time span of the dataset increases. Mentioned compulsory eliminations are the reason of this decrease in used HARPs. This is also the reason why the machine learning algorithms were unable to perform better in larger dataset which is unique to this problem in this study whereas, usually more data results in better results. Moreover, since SHARP data is in 12-minute intervals. So, to label the CME classes onto the SHARP data during preparation of datasets the CMEs from the CDAW database were shifted to multiples of 12 in the minutes section wherever it was necessary. Otherwise, there are cases where the CME lands in between two different rows of HARP data. So, in turn during the calculation of metrics during testing the performance of the models some predictions with a 12-minute difference with original data were accepted as correct. A depiction of the datasets and their complexity along with pulse generated is given in figures 3 till 5 and the appendix section. An increase in complexity towards larger datasets can be observed from these figures.

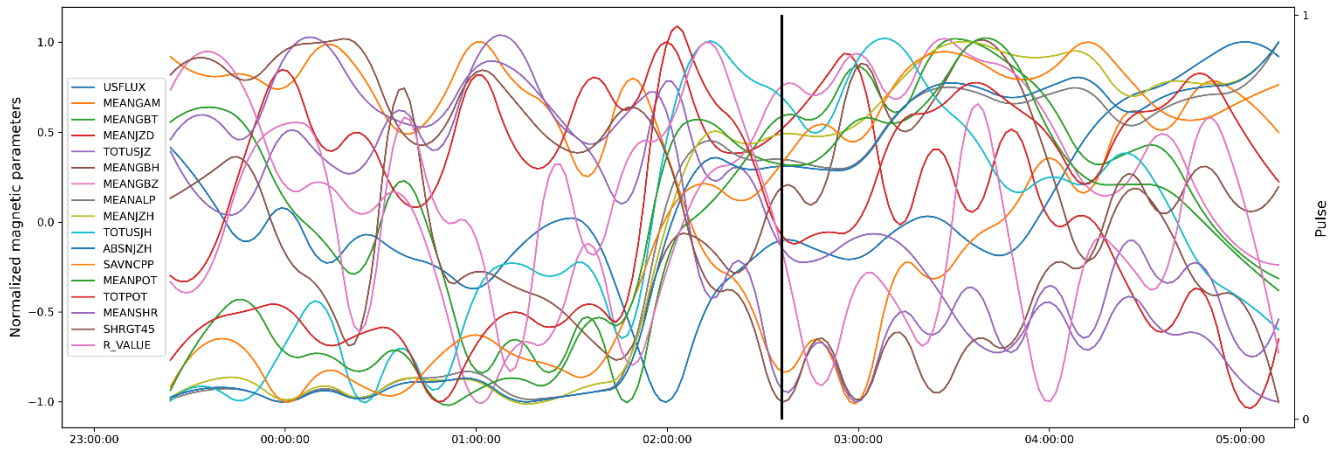



Figure 4. A representation of the first HARP of the 6-hour dataset along with the pulse parameter generated for it based on CME timing from CDAW database.

### 3.2.5 Conversion Of The Datasets In A Timeseries Format For Machine Learning Algorithms

Machine learning algorithms used in this study especially LSTM uses window structure to simultaneously observe changes in time-dependant data. This window size is set to 1 hour for all the machine learning algorithms as this was found to be most effective. When feeding the networks, the data samples are created by sliding the window forward. The window (e.g., will span 5 data points for 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. The label for the data where there is not a CME at the last entry meaning the last timestep would be 'NOTCME' whereas the label for the data where the last data has a pulse value of 1 would be 'CME'. So, in fact it answers the question: 'Does the change in this set of data points result in a CME?' The window will start from the beginning of a HARP till the end. Then, move to a next HARP in the dataset. The predictions from 2011 to 2020 were also made by providing the data in question in this form. E.g., the HARP was divided into such samples and label (pulse parameter) was predicted by the models for each sample.

Table 3

The datasets developed for each time span and then used for the machine learning models.

| 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 |

The length of the datasets given in the second column of table 3 shows the length of the different datasets prepared in the form of matrix where the first entry in this matrix shows the number of datapoints in the dataset. The second entry in the matrix shows the window size which the moved across the length of the datasets consisting of different number of HARPs while labelling and the producing the pulse of the dataset to depict the presence or absence of CME at that point in the time-series. This also is the number of the data point for the 17 different time-series that the models are exposed to simultaneously while training and prediction. The third entry in the dataset shows the number of time-series in the dataset which is the number of 17 different magnetic parameters of HARPs. This depiction shows the length of the dataset as well as the complexity of the dataset at the same time.

The output classes are categorized to form two different classes for presence and absence of CME, respectively instead of a single class. This is usually done to force the artificial intelligence models to learn the characteristics of both cases i.e., the presence and absence of CMEs treating them as two different events and paying equal attention to both cases. Otherwise, the problem of over fitting becomes unavoidable as has been tried and ratified in the course of this study.

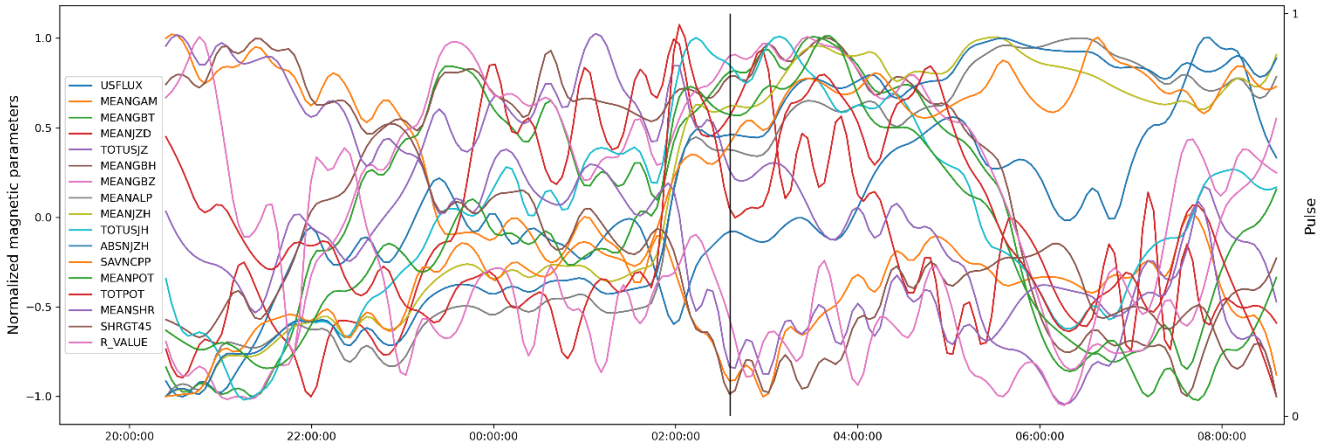



Figure 5. A representation of the first HARP of the 12-hour dataset along with the pulse parameter generated for it based on CME timing from CDAW database.

### 3.2.6 Implementing Weights To Each Class To Overcome Biased Learning Of ML Models

Table 4 shows the distribution of the two classes in the datasets along with the percentage of the dominant class in each dataset. This shows the imbalance of the CME and NOTCME class in each dataset. This is natural as there are more times where there is no CME as compared to when there is a CME event. This situation is very unhealthy to train any AI model because this means that the model eventually labels every data it is given as NOTCME and still get accuracies upwards of 95%.

Table 4

The number CME events versus the total events in different datasets and their imbalance ratios.

| Dataset | Number CME events | Total events in dataset | Imbalance percentage |
|---|---|---|---|
| 2-hour dataset | 78 | 780 | 10.00 |
| 4-hour dataset | 59 | 1200 | 4.91 |
| 6-hour dataset | 44 | 1320 | 3.33 |
| 8-hour dataset | 31 | 1240 | 2.50 |
| 10-hour dataset | 23 | 1150 | 2.00 |
| 12-hour dataset | 20 | 1200 | 1.66 |

The models will never try to learn the difference between the wo classes as this will cause decrease in performance. As can be seen from the last column of Table 4 that the imbalance is abnormally huge. In the case of the 8-hour dataset, for example, if the model predicts all the data as NOTCME realistically it will achieve an accuracy of 97.5%. To overcome this problem weights were added to each class depending upon the frequency of that class in the dataset. This means that the machine learning models are rewarded differently for correctly predicting both classes. Similarly, the models are penalized differently when they predict NOTCME class wrong than when they predict CME class wrong. This means correctly predicting CME class will yield more reward than NOTCME class and predicting the CME class wrong will result in bigger penalty than predicting NOTCME class wrong. This ensures that the model try to learn the difference between the two classes as much as possible and try to minimize the error while training. The following function was used to add weights to the classes to overcome imbalance.

```
weight_for_0 = (1 / num_of_notcme)*(len(data_all))/2.0
weight_for_1 = (1 /  num_of_cme)*(len(data_all))/2.0
class_weight = {0: weight_for_0, 1: weight_for_1}
```

The piece of code shown calculates the weights so that the impact of each class is approximates to 50% in each dataset.

### 3.2.7 Division Of Training, Validation, And Test Datasets For Training

The datasets were divided into training, validation and test datasets for training and testing the models. For models based on Decision Tree, kNN, SVM and Random Forest algorithms 75% of the datasets were used as training datasets and used for training the models and 25% of the datasets were used as test datasets and used for the evaluation of the performance of the models. For LSTM models 70% of the datasets were used as training dataset used for training the models, 15% of the datasets were used as validation datasets used during training for boosting the training performance and 15% of the datasets were used as test dataset used for the evaluation of the performance of the LSTM models. This data was never shown to the models during the training procedure and only used for evaluation of the models. The datasets were shuffled before division. The shuffling is performed in terms of complete HARPs to avoid breaking the time-series.

Table 5

An illustration of the division of dataset into training, validation, and test datasets. Here lengths of datasets created from the 4-hour dataset are shown as an example.

| 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) |

These shuffled time-series data of the 17 parameters of HARPs along with their pulse was then divided into training, validation, and test datasets. This division is performed without breaking the series created by the sliding the window of time span 1 hour on the time-series separately. A new pulse is generated at this step so that pulse reflects the changes in each sample with respect to presence or absence of CME for the given change. One-hot encoding is performed on the pulse parameter at this stage to convert the pulse parameter in categorized classes. CME class and NOTCME class, respectively. This helps in the performance of the recurrent networks like RNNs and LSTMs. The following piece of code divides the data into training, validation, and test datasets.

```
def train_test_split(X, Y, **options):

    t_v_size = options.pop('test_vald_size', None)
    if t_v_size is None:
        t_v_size = 0.16
    n_sample = len(Y)
    n_vald = int(n_sample * (t_v_size/2))
    n_test = int(n_sample * t_v_size)

    X_train = X[:-n_test]
    Y_train = Y[:-n_test]

    X_test = X[-n_vald:]
    Y_test = Y[-n_vald:]

    X_vald  = X[-n_test:-n_vald]
    Y_vald  = Y[-n_test:-n_vald]
    return X_train, Y_train, X_vald, Y_vald, X_test, Y_test, n_vald
window_size = 5
X_all, Y_all = create_dataset(X, Y, look_back=window_size)
X_train, Y_train, X_vald, Y_vald, X_test, Y_test, n_test=train_test_split(X_all,
Y_all, test_vald_size=0.30)
```

The window size here is given as 5 because in our dataset the cadence of data is 12 minutes and 5 reflects one hour which is the selected window size for the ML algorithms.

### 3.2.8 Creation Of Decision Tree, SVM, Knn And Random Forest

In the favour of easier comprehension of complexity of the problem at hand, each dataset i.e., 2-hour dataset, 4-hour dataset, 6-hour dataset, 8-hour dataset, 10-hour dataset and 12-hour dataset includes 17 different time-series' which reflect the 17 different magnetic field parameters of the source active regions. An additional parameter which reflects the presence or absence of CME. The ML-algorithms/deep-learning algorithms have to learn or find patterns in these 17 time-series simultaneously and successfully predict the pulse parameter which reflect the presence or absence of CME. This explains the difficulty of this task if done manually and also the scarcity of the identification of source active regions. Learning the magnitude of the changes and the effect causes by these changes in the magnetic field parameters is essential to correctly predict source active regions and hence deploying machine-learning/deep-learning is efficient and necessary (Bobra & Ilonides, 2016). Deep-learning techniques especially LSTM are very successful and efficient in catching patterns in time-series data (Inceoglu et al., 2018; Florios et al., 2018).

In the creation of Decision Tree (DT), the minimum number of instances in leaves was set to 95. The minimum split subset was found to be most efficient at 11. Optimal results were obtained when the maximal tree depth was set to 4, respectively. The binary tree was included in the training procedure. The training of DT was stopped when the majority reached 90%. The structure of DT formed after the training process for 2-hour dataset is shown in Figure 6. The splits at level 4 have already reached at 100% accuracy but there is a split 100% accuracy on the other side of split. The results show poor results on test dataset as compared to LSTM network. Further results are discussed in the results section. Unsurprisingly, LSTM network proved to be the best performing algorithm in this analysis further discussed in the next section.

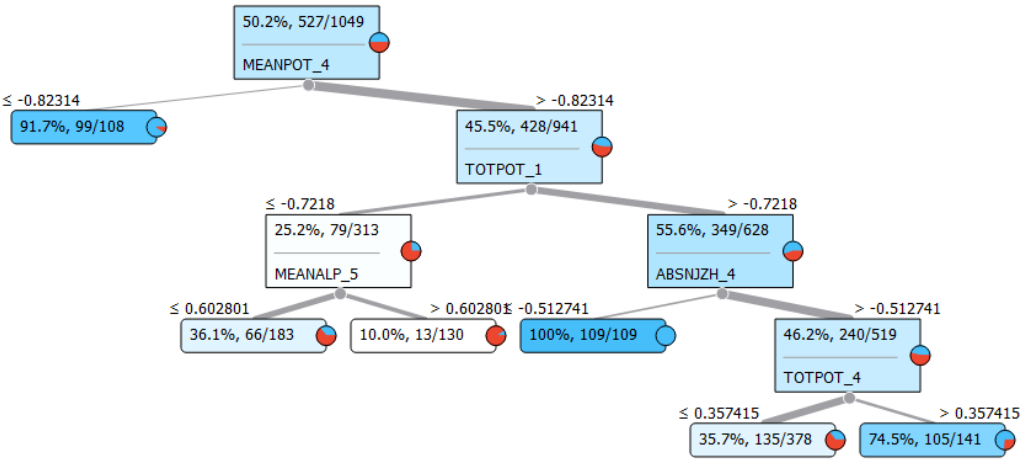



Figure 6. An illustration of the formed DT on 2-hour dataset. Training dataset is used during training and model is tested on test data and results are presented in results section.

For SVM models optimal results were obtained when the cost was set at 1 and the iteration was limited to 100. Polynomial and sigmoid kernels gave poorer results, so these were omitted during the training procedure for SVM. The linear kernel gave the best results, this was set along with a 10% regression loss epsilon to maximize the performance of the models. The number of neighbours were set to 17 with the weight of the models being distance with Manhattan metric. kNN usually performed very poorly in our analysis. For RF, the number of trees were optimized at 10 with a maximum of 5 attributes considered at each split. Replicable training was performed on models formed with RF. The depth of each tree was limited to 3 and the maximum subsets were limited to 5 as performance of the RF models suffered after more splits.

### 3.2.9 LSTM Models And Their Hyperparameters

LSTM has proven repeatedly to be very efficient at learning patterns in time-series data and sequential data like speech recognition and text-to-speech applications (Karim et al. 2019). There are four main parts to a LSTM unit. This unit can be considered equivalent of neuron in a traditional neural network. The LSTM unit contains a memory cell which is the main reason these networks are so efficient in pattern recognition, more on this later, a forget gate, an input

gate, and an output gate. The input and output gates permit the data in or out of the cells. These gates control what amount of data under what conditions should be allowed in or out of the LSTM cell. The memory cell holds certain information in them which help them take previous information in consideration while training with respect to new information. The forget gate controls which information contribute less or nothing to successful predictions and thus can be lost with any cost to performance. Figure 7 shows an illustration of the LSTM unit as used in this study.

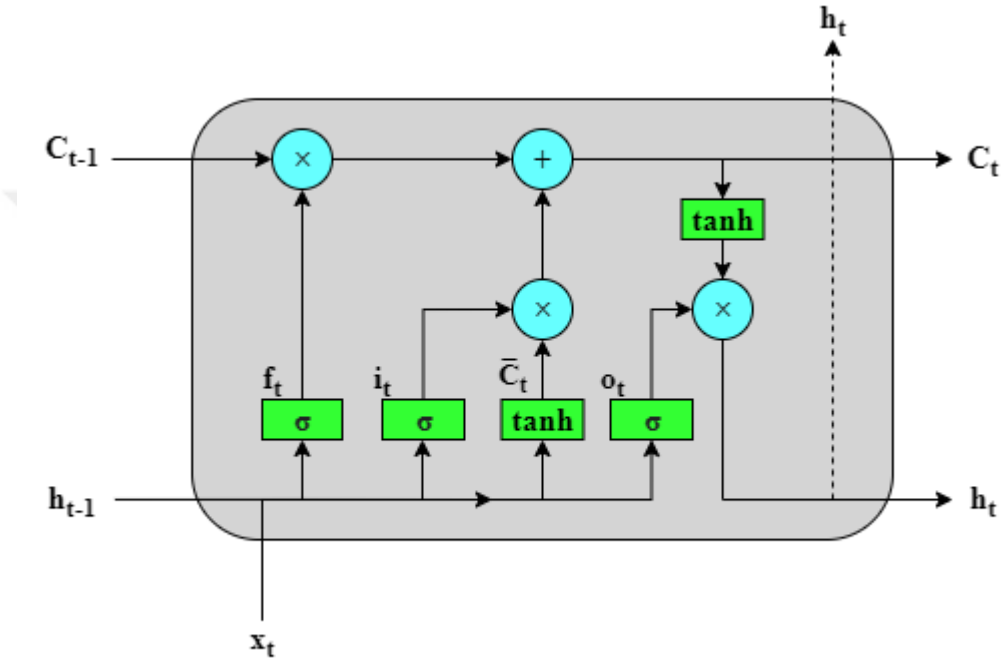



Figure 7. A depiction of LSTM unit. Here$X_t$ represents input vector given to the LSTM unit, $h_t$ shows the output vector produced by the LSTM unit, whereas $f_i$, $i_t$, $o_t$, $c_t$ represent the forget gate, input gate, output gate and the cell state, respectively.

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t \quad (1)$$

The old state shown by $C_{t-1}$ is updated by the new state $C_t$ which is the candidate state and is calculated by the expression given above (1). The forget gate $f_t$ controls the amount of data to be remain within the cell. The values of forget gates are calculated using equation 2.

$$f_t = \sigma\left(W_f \,.\, [h_{t-1}, x_t] + B_f\right) \quad (2)$$

The input gate $i_t$ controls the amount of the data to be preserved in the cell or to be lost to increase performance. Equation 3 determines the value of the input gate. $X_t$ in equation 3 depicts the input vector at time step t and $h_{t-1}$ is the output vector at time step *t-1*.

$$i_t = \sigma(W_i \,.\, [h_{t-1}, x_t] + B_i) \quad (3)$$

The candidate cell state can be calculated by equation 4 and the is represented by *Ct.* This shows the shape and value of the candidate vector.

$$C_t = tanh(W_c \,.\, [h_{t-1}, x_t] + B_c) \quad (4)$$

To calculate the output vector at time step *t*, which is represented by $h_t$, equation 5 can be used. The variable $o_t$ in equation 5 can be calculated by expression given in equation 6.

$$h_t = o_t \odot \tanh(C_t) \quad (5)$$

$$o_t = \sigma(W_o \,.\, [h_{t-1}, x_t] + B_o) \quad (6)$$

The parameters *W* and *B* in above equations represent the weights and bias of the LSTM model, respectively. These parameters have to learned during the training of the models. In the above equation (·) represents the concatenation of the vectors involved in the equation, where $\sigma(\cdot)$ represents the sigmoid function, $tanh(\cdot)$ represents the hyperbolic tangent function and (·) represents the Hadamard product. The sigmoid, hyperbolic tangent functions can be expressed as equation 7 and 8, respectively.

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (7)$$

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (8)$$

The LSTM network is trained by giving all the data in the window size to the model simultaneously and the corresponding target variable, that is, the pulse is then targeted for training. The weights and bias of the models are then optimized to reach maximum correct target variable with respect to the weights assigned to the class explained in section 3.2.7.

The LSTM models were trained using batch size as 1. This was done due to maximize the training potential of the models. A dropout amount of 0.2 was set during the training process of the LSTM models. This setting is used to minimize the overfitting of the models. Overfitting can be described as a condition during the training process where the models have seen all the data samples and learn the individual values of the target class to the data samples introduced to the models and thus perform near perfect on the training data but very poorly on the test data which has not been seen by the models. This dropout keeps some of the data from the data samples out of the training process within each iteration. This decreases the chances of overfitting. The epoch number of the LSTM was set to unlimited, and a stopping criterion was

introduced during the training process, this stops the training by monitoring a selected metric during the training continuously and thus the number of epochs can be set to unlimited, and the model stops the training when the improvement of the performance seizes. The accuracy with the validation data was used as the metric for early stopping criterion. A patience parameter was set, and the value was set to 5 to trigger the early stopping criterion.

The number of neurons set for each model is different and was selected by optimizing the performance of the models. For example, the number of the neurons is 50 in the LSTM layer of model designed for 4-hour dataset whereas, the number of neurons in the LSTM layer for 10-hour dataset is 150. These parameter changes for each dataset as the models perform differently with different hyperparameters for different datasets depending on the length of the dataset. Tanh activation function is used in the LSTM layer of the models. The loss function used in the LSTM models is the categorical cross entropy with logits activated in the layers. This loss function is specifically designed for categorical data as target class and performs very well with categorical target data (Yendapalli et al., 2020). Adaptive Moment Estimation also known in Adam in deep-learning community is used as the optimizer in the LSTM models (King & Ba 2015). This is actually a technique used as a method for stochastic gradient decent. The results produced by the LSTM models and other models based on DT, SVM, kNN and RF are given in the result section of this thesis.

### 3.2.10 Filtration Of HARPs Before Performing Identification Of Source HARPs

After the production of a reliable LSTM model which is the model trained on the 4-hour dataset, before the deployment of the model to perform identification of source HARPs for CMEs in CDAW from 2011 till 2020, a filtration process was employed on all HARPs to make an AR-Pool for each CMEs containing all possible candidate source HARPs. This helps us maximize the performance of the model as HARPs which are impossible to be source HARP for a particular CME are eliminated. This filtration process can be defined as follows.

For a particular CME:

I All HARPs that are present on the solar surface at the same exact time are selected. Compensation for the cadence of 12 minutes is done beforehand as explained in section 3.2.4.

II Within the selected HARPs for this CME, HARPs with a position angle of ±90$^{o}$ are selected. This is because the POV of CME is so that any HARPs outside of this hemisphere i.e., position angle within 90$^{o}$ are impossible to initiate that CME.

III Within the HARPs selected in step II, only those which have spent at least 4 hours in the area designated in step II are selected and added to an AR-Pool for this CME.

The third step of this filter is designed so that any HARPs that happen to be inside the designated area on the solar surface of that CME that is ±90$^{o}$ of that CME but are at the edges of solar surface are eliminated. This elimination is important because these HARPs continue to advance to the other side of the Sun or travel to an area from these cannot be involved in the onset of the CME in question.

After the filtration process is performed on all the HARPs from the SHARP database with respect to CME from CDAW database an AR-Pool is formed for each CME. The HARPs in the AR-Pool for each CME are subjected to the LSTM model used for identification procedure and the model ideally selects a single HARP for each CME as the source HARP for that CME. A database formed by this procedure along with the models and data used can be reached at Raheem et al. (2021), which is the data repository for this thesis. The results produced regarding the filtration process and then prediction procedure of LSTM model to perform the predictions are given in the result section.

# CHAPTER 4
# RESULTS AND DISCUSSION

## 4.1 Performance Metrics Deployed To Evaluate ML Algorithm-Based Models

To evaluate the performance of different ML algorithms used in this study precision, recall, F1 score and Accuracy metrics were used. Precision, recall and F1 score are widely used for AI models that are used for classification purposes. Precision can be defined as a metrics that checks what ratio of identification that are identified as positive, are actually positive in the dataset. This can be shown by equation 9 given below.

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

On the contrary recall can be defined as a metric that checks what ratio of the positive labels in the dataset were identified correctly by the model. This is calculated by equation (10)

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

The expression TP, TN, FP, FN in the equations shown here represent true positives, true negatives, false positives, and false negatives, respectively. These show the number of correct positive identifications, the number of correct negative identification, the number of incorrect positive identification and number of incorrect negative identification, respectively.

The metric F1 score shows the balance of recall and precision metric. This represents the overall ability of the model to generalize the problem. F1 score can be calculated by equation 11.

$$F_1 = 2.\frac{precision\,.\,recall}{precision+recall} \tag{11}$$

Accuracy of the models can be defined as a ratio of correct predictions performed by a model to the total number of the predictions made by the model. Accuracy can be given as expression given in equation 12.#

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

These metrics were calculated for each model during and after training processes to evaluate the performance of the models within themselves and with other models to determine the best technique for the problem tackled in this thesis.

### 4.2 Results For ML Models Based On DT, SVM, Knn And RF Algorithms

Five ML algorithms were deployed in this study i.e., DT, SVM, kNN, RF and LSTM to find the best model to perform the identification procedure for source HARPs for CME from 2011 till 2021. The results show generally a poor performance by DT, SVM, kNN, and RF as compared to LSTM models. This result is not surprising as the problem at hand as discussed in section 3.2.8 is quite compilated which includes multiple time-series data with a single data sample that have to be considered simultaneously to learn the changes and/or patterns in the data to perform the identification and learn the changes involved behind the onset of CMEs.

Table 6

The performance metrics for the models based on DT, SVM, kNN and RF on all the datasets used in this study.

| | 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 6 shows the results metrics for DT, SVM, kNN and RF. Note that these results are subpar and none of these models are able to score an accuracy higher than 65%, a precision of 64% and a recall of 65%. These results depict both the complexity of this problem along with the incapability of these algorithms to perform on time-series data. Table 7 shows the confusion matrix for these models, and it can be further understood that a more sophisticated algorithm was necessary in this study.

Table 7

The confusion matrix for the models created using DT, SVM, kNN and RF algorithms.

| **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% |

### 4.3 Results Produced By LSTM Networks On Different Datasets

Results produced by the LSTM networks both during the training procedure and testing procedure are shown in table 8 and table 9. These show a great contrast between the results shown above and here.

Table 8

The performance of LSTM models trained on different datasets during training procedure.

| **Metric** | | **Size of the training data** | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2-hour | 4-hour | 6-hour | 8-hour | 10-hour | 12-hour |
| Accuracy | Training | 0.964 | 0.97 | 0.954 | 0.82 | 0.91 | 0.87 |
| | Validation | 0.872 | 0.9 | 0.859 | 0.72 | 0.94 | 0.85 |
| | Test | 0.803 | 0.863 | 0.878 | 0.77 | 0.71 | 0.81 |
| | All | 0.937 | 0.944 | 0.928 | 0.8 | 0.89 | 0.87 |

The performance of the LSTM model is superior to the previous models and thus LSTM was considered the best technique for the task at hand. Figure 8 shows the loss of the selected model i.e., the LSTM model trained on the 4-hour dataset which is also the best performing model as depicted by table 9. with respect to epochs during the training process.

Table 9

The performance of LSTM models trained on different datasets after the training procedure. The performance shown here is evaluated on test dataset.

| Result Metric | Size of the datasets | | | | | |
|---|---|---|---|---|---|---|
| | **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 |

The selected model was then used later on during the identification process for source HARPs. Boldface is used in table 9 to show the performance of this model trained on 4-hour data. The precision, recall, accuracy and $F_1$ score of this selected model is 81%, 77%, 86% and 79%, respectively. The model is well suited to be used for prediction process as shown by the results. No model during the study was able to perform better than this model. Performance of the models trained on larger datasets i.e., 10-hour, 12-hour datasets is poor contrary to normally what could have been expected. The reason for this can be the lack of sufficient data available in these datasets as when the HARPs that cannot be correctly labelled by checking CDAW database are removed the number of HARPs used in these larger datasets decreases dramatically. This is shown in section 3.2.5 table 3.

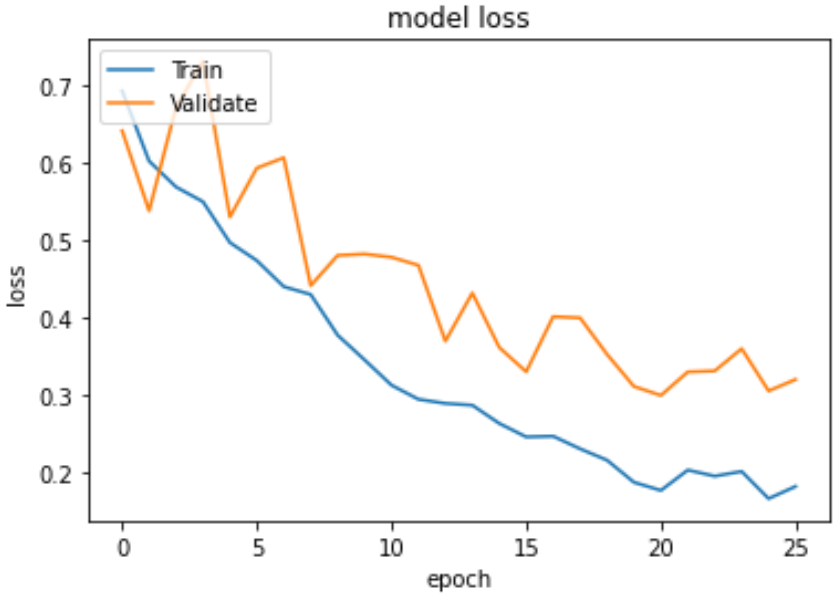



Figure 8. The change in the loss of the LSTM model trained on 4-hour dataset with respect to epoch.

Figure 9 shows an illustration of the performance with respect to predictions made by the selected LSTM model on training, validation, and test dataset all together for the 4-hour dataset. Here the actual dataset is shown in (a), training dataset, validation, and test datasets are shown in blue, black, red, respectively. (b) shows the predictions made by the selected LSTM model.

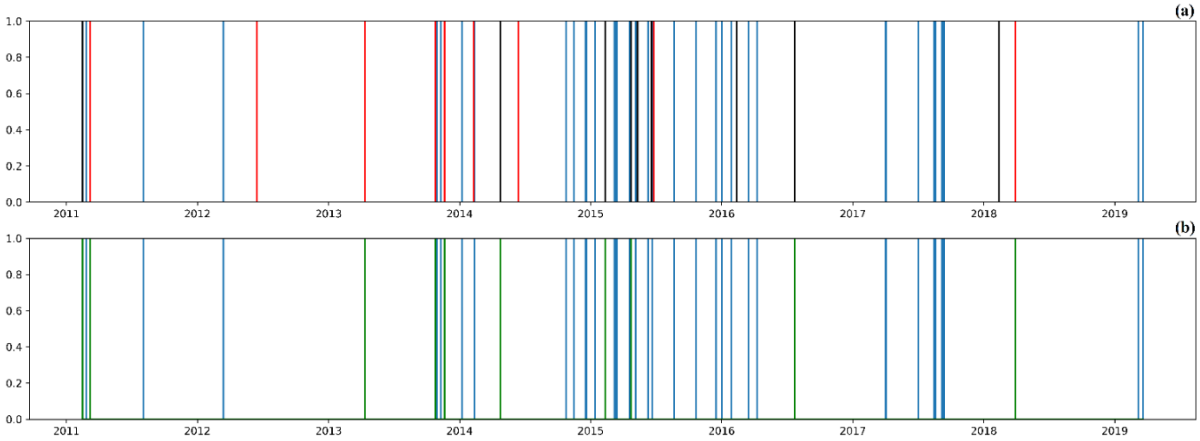



Figure 9. The performance of the selected LSTM model on training, validation, and test datasets altogether. (a) shows the actual dataset whereas, (b) shows the predictions made the LSTM model with respect to years. Here each vertical line represents a single CME in the 4-hour dataset.

The figure 9 was designed to highlight the following points regarding the selected LSTM model and the predictions made by the model.

I It shows the distribution of training, validation and test datasets within a dataset used. It shows the shuffling of datasets before these mentioned datasets to remove any bias.

II It shows that the dataset in very random as the HARPs used are from different years spanning 10 years. Also, that they are not equally distributed over the years so there no periodicity present with respect to their position in time. So, the model cannot just periodically predict presence of CMEs without considering the magnetic parameters.

III It simultaneously shows the performance of the model over all sets namely, training, validation, and test i.e., during training and during testing.

### 4.4 Identification of source HARPs using the selected LSTM model

After the training and before the deployment of this model on the magnetic field data for HARPs from SHARP, the HARPs were subjected to the filtration process as discussed in section 3.2.10. This filtration process produces an AR-Pool for each CME from 2011 till 2020 present in the CDAW database. Table 10 shows the statistics obtained after the filtration process.

Table 10

Results of the filtration process discussed in section 3.2.10 performed before the deployment of the LSTM model to perform the identification procedure.

| **Year** | **CMEs** | |
|---|---|---|
| | Total CMEs listed in CDAW | CMEs 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 |

As a lot of CMEs have an AR-Pool, it can be said a very little portion of CMEs are initiated at a time and an area where no HARPs are visible. These CMEs could be initiated by HARPs present on the other side of the soar surface which is not observed by SDO spacecraft. The number of CMEs for which a single HARP could be selected from the AR-Pool after subject to the LSTM model decreases as a lot of CMEs are initiated by HARPs present on the unobserved side of the Sun. Table 11 shows the results obtained after the CME with their AR-Pool are subjected to the LSTM model to identify the source HARP based on the learning done during training procedure.

Table 11

Results of the identification process performed after the deployment of the LSTM model to perform the identification procedure.

| 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 |

The LSTM model is given the magnetic field data of the HARPs within an AR-Pool at a time in the same format as the dataset used during training procedure i.e., [+1,-1] normalization with respect to each HARP is performed, padding is added where there no data available, the data is converted into same time-series format as during training procedure. The LSTM model predicts the pulse parameter for the given input data. This pulse is then converted back to correct time format and then CDAW database is searched for the predicted CME. If there is a CME at same time as the prediction and the prediction of CME is performed only on a single HARP from the AR-Pool the prediction is considered correct and added the database produced as a result of this study also available at Raheem et al., (2021) (consult last column of table 11 for these results).

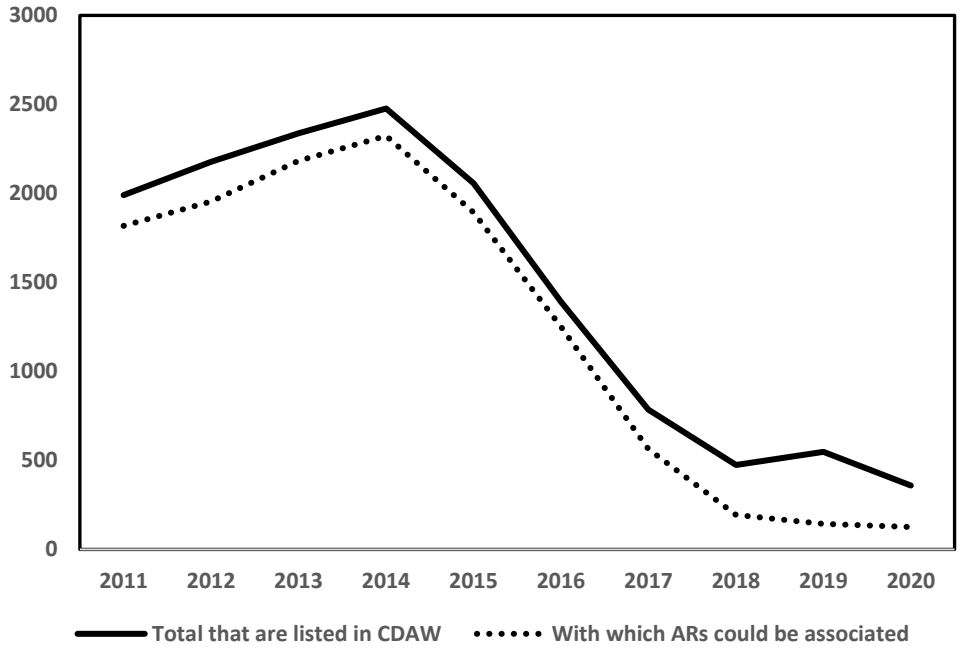



Figure 10 The change in the total number of CMEs present in the CDAW database with respect to the CMEs where an AR-Pool was formed after the filtration process was performed on these CMEs mention in section 3.2.10.

The figure 10 shows the variation in the number of CMEs for which an AR-Pool with HARPs from the SHARP database whereas figure 11 shows the CMEs for which a single HARP was selected from the formed AR-Pool by the LSTM model used for prediction during the identification procedure. These findings increase the number of the identified source HARPs for known CME in the CDAW database by 31 folds as the number of CMEs with identified source active regions is a mere 156 in the DONKI database and the number of CME with identified source active as a result of this study is 4895 (Raheem et al., 2021).

If we take a statistical look at the results presented in table 11, 4895 CMEs have been associated with source active region out of the total 14604 CMEs in the CDAW database. This makes a total of 33.5% of the CMEs for s source active region has been identified. If we consider the field of view (FOV) of the SDO spacecraft from which the data is obtained and used in this study, this is an excellent increase in the number of CMEs with an identified source active region. If the other side of the Sun is also considered, then the number of identifications made in this study can be considered to be around 67% of the total CMEs. If those HARPS

were also included in the training dataset this 67% of identification could have been made using the same LSTM model produced in this study.

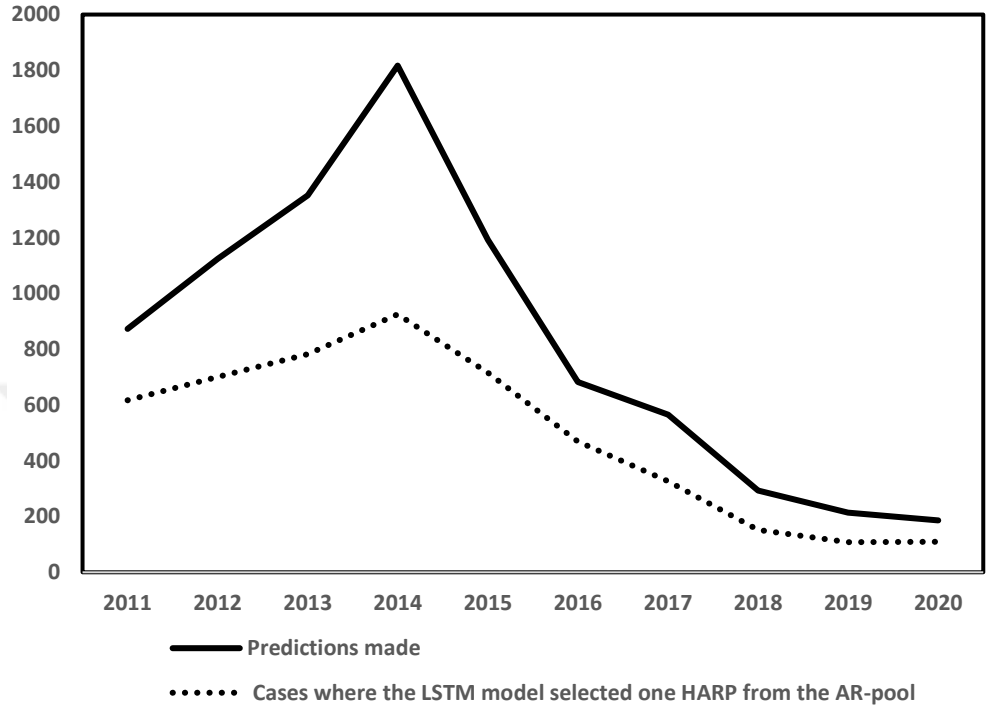



Figure 11. The change in the total predictions made by the selected LSTM model during the identification procedure with respect to the CMEs for which a single HARP was selected as source HARP for that CME by the selected LSTM model as shown in Table 11.

### 4.5 Analysis Performed On The Magnetic Field Parameters Of İdentified Harps With Respect To Their Respective CMEs

An analysis is performed, as the relationship between the magnetic field parameters of HARPs as mentioned in Table 11. with respect to their respective CMEs as identified in the course of this study. Especially, current related parameters namely, MEANJZD, TOTUSJZ, SAVNCPP of the identified HARPs are their change with respect to the onset of CMEs and the properties of CMEs namely, the linear speed, acceleration, mass, and KE of the CMEs are investigated. These show some interesting and note-worthy findings. These magnetic properties of the source HARPs are important as these are motion related. These findings are given in the points listed below. Figure 12 shows these changes in graphical manner. The values of the magnetic field parameters of the source HARPs and CME parameters were normalized

between 0 and 1 to compensate for the large differences in the scales of these parameters. The values for MEANJZD for source HARPs and the acceleration of CMEs were normalized between -1 and +1 to adjust for the negative values in the original data.

I The linear speeds of the CMEs initiated by the source active regions identified in this study are all below 1058.54 km/s mark. The masses these initiated CMEs are below $1.33e^{+16}$ whereas, the energies of these CMEs are around $6.20e^{+31}$ erg on average (consult Figure 12 (a) through 12 (l) for detail on this).

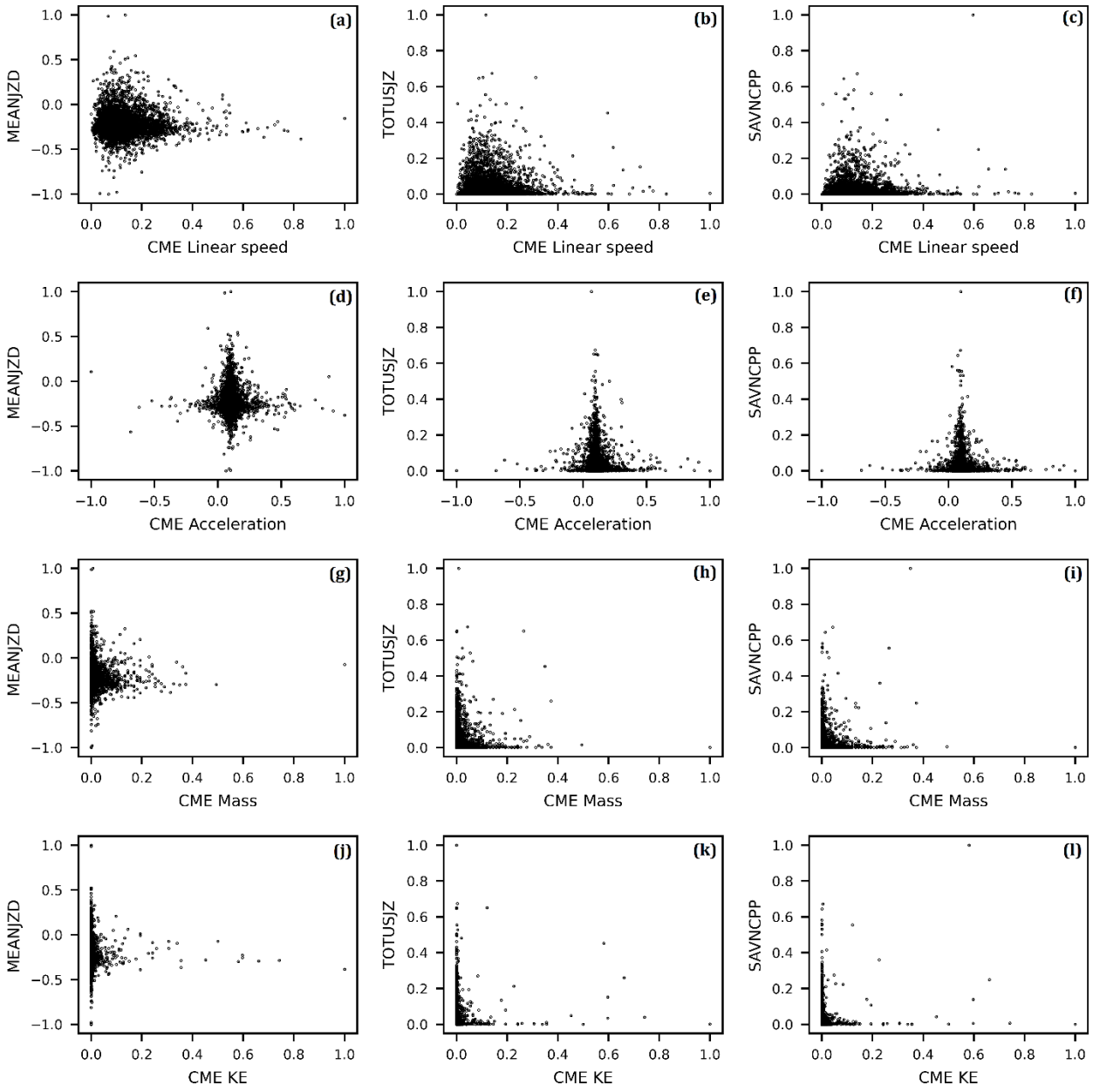


Figure 12. The relationship between the magnetic field parameters of source active regions i.e., MEANJZD, TOTUSJZ and SAVNCPP and the properties of CMEs initiated by them.

II The MEANJZD also known as mean vertical current density parameter for all the HARPs listed in SHARP database changes from a minimum of -4.39 $mA/m^2$ and reaches to a maximum of 7.69 $mA/m^2$. The minimum and maximum of the mean vertical current density of the HARPs identified as source for particular CMEs in this study are -1.39 $mA/m^2$ and 2.86 $mA/m^2$, respectively. This infers that only solar active regions with mean vertical current densities between this range seem to play a role in the initiation of CMEs (consult figure 12).

III Similar to point II the value of parameter TOTUSJZ in SHARP for HARPs is around $7.6e^{+12}$ A whereas the value for total unsigned vertical current for the source HARPs identified in this study is around $6.5e^{+13}$. The average value for sum of the absolute value of the net current per polarity denoted by SAVNCPP parameter for identified source HARPs is approximately around $3e^{+13}$A (consult figure 12).

IV The total unsigned vertical current (TOTUSJZ) of the source HARPs and the kinetic energies of the CMEs initiated by them are inversely proportional and have an asymptotic relation with each other. The same relationship is present between the kinetic energies of CMEs and the sum of the absolute value of the net current per polarity of their source HARPs.

V At average, the total unsigned vertical current, and the sum of the absolute value of the net current per polarity of the source HARP is inversely proportional to the mass of the associated CME.

VI Source HARPs with negative mean vertical current initiate accelerating CMEs according to the analysis performed on the source active regions identified in this study (See, Figure 12 (a,d,g,j)).

VII 84.42% of the CMEs in the database are below the average value for the momentum of the CMEs for which a source HARP could be associated. The average value for momentum for these identified CMEs is $7.88e^{+17}$ gkm/s (See, Figure 12).

These results provide important insight to understand the mechanism behind the onset of CMEs. These results should be taken critically since this is the first attempt at such a large-scale identification of source active regions for CMEs. The mechanism developed in this study can absolutely benefit from an increase in the training data.

The next section goes over the conclusions reached in this thesis and the validity of results obtained through this proposed mechanism.

# CHAPTER 5
# CONCLUSION

The mechanism proposed in this thesis brings a unique new technique to automate the identification process of source active regions for CMEs, a process which is done manually until this point as to my knowledge. 4895 source active regions have been identified for CMEs using this mechanism in the course of this study. These identified source active regions can boost the studies regarding the understanding of mechanisms behind the onset of CMEs. This study is unique and one of a kind in this regard. The quality of the identifications made in this study using the LSTM model can be increased by an increase in the data available to train the deep-learning models as the LSTM model was trained on a very limited data available for the magnetic field parameters of the source active regions. The certainty of the data used for training is not quantified and the amount of noise and other instrument related errors are present in the data. However, the model has been successful in regard of the precision on the test dataset and the results produced. The precision was intentionally paid a great importance while training as the quality of the predictions is desired over the quantity of the predictions. The number of correct predictions was more important than the number of total predictions made by the model. The confidence level of the predictions is, therefore, kept high as the model with highest overall precision was eventually selected to perform the identifications.

Moreover, the correlations between the CME events for which an AP-Pool could be formed and the total CMEs listed in the CDAW database with respect to years is 99.8%. Whereas the correlation between the number of the total predictions made by the selected LSTM model and the events where only one single HARP was selected from the AR-Pool made for the CMEs is 97.9% over the years. This shows an absence of random selections by the LSTM network. These are shown in Figures 10 and 11. This further demonstrates the ability of the selected LSTM model to generalize the problem and that the model is not making random predictions. The correlations of the results based on the predictions made by the LSTM network which are added to the database formed and where one HARP was selected by the model from the AR-Pool, with sunspot numbers from the American Association of Variable Star Observers (AAVSO) and the Sunspot Index and Long-term Solar Observations (SILSO) is 98.5% and 98.1%, respectively.

Table 12

The correlations between the sunspot numbers obtained from AAVSO and SILSO and the CMEs for which source active regions are identified in this thesis.

| 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 |

This highlights that the predictions made the LSTM model are also consistent with the solar cycle. This can also be observed from figures 10 and 11, as although the solar cycle is not graphed in these figures, a solar cycle can be seen, and 2014 solar maximum is visible in these figures (Consult y-axis of figure 11). This is because of the positive correlation with the solar cycle.

If the correlations between the parameters of source active regions and the parameters of the CMEs initiated by them is examined, all correlations are above 80% with the highest being 96%. These correlations are given in table 13. All values here are aggregated annually for each year before checking the correlations. This further shows the accuracy of the predictions as these results would not have been this high if the selected LSTM model had produced random predictions in the identification process.

Table 13 The correlations between the parameters of source active regions and the parameters of the CMEs initiated by them identified in this study as a function of annual aggregation for each year.

| CME Parameters | Source Active Region Parameters | | |
|---|---|---|---|
| | MEANJZD | TOTUSJZ | SAVNCPP |
| Linear speed | 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 |

This technique has produced 4895 identifications of source active regions for CMEs whereas the DONKI database only listed 120 usable and 156 total source active regions. That is an increase of 31 folds over the previous identifications. This technique provides a unique opportunity to boost the studies regarding investigation of the mechanisms behind the onset of CMEs and space-weather studies in general.

This study has been published in Monthly Notices of the Royal Astronomical Society (MNRAS) and can be accessed from Raheem et al. (2021).

## REFERENCES


Alex, S., Mukherjee, S., & Lakhina, G. S. (2006). Geomagnetic signatures during the intense geomagnetic storms of 29 October and 20 November 2003. *Journal of Atmospheric and Solar-Terrestrial Physics*, 68(7), 769–780. https://doi.org/10.1016/j.jastp.2006.01.003

Berkebile-Stoiser, S., Veronig, A. M., Bein, B. M., & Temmer, M. (2012). Relation between the coronal mass ejection acceleration and the non-thermal flare characteristics. *Astrophysical Journal*, 753(1), 88. https://doi.org/10.1088/0004-637X/753/1/88

Bobra, M. G., & Ilonidis, S. (2016). Predicting coronal mass ejections using machine learning methods. *The Astrophysical Journal*, 821(2), 127. https://doi.org/10.3847/0004-637x/821/2/127

Bobra, M. G., Sun, X., Hoeksema, J. T., Turmon, M., Liu, Y., Hayashi, K., … Leka, K. D. (2014). The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs - Space-Weather HMI Active Region Patches. *Solar Physics*, 289(9), 3549–3578. https://doi.org/10.1007/s11207-014-0529-3

Cavus, H., Araz, G., Coban, G. C., Raheem, A. ur, & Karafistan, A. I. (2020). Correlation between sunspots and interplanetary shocks measured by ACE during 1998–2014 and some estimations for the 22nd solar cycle and the years between 2015 and 2018 with artificial neural network using the Cavus 2013 model. *Advances in Space Research*, 65(3), 1035–1047. https://doi.org/10.1016/j.asr.2019.09.056

Falconer, D. A., Moore, R. L., & Gary, G. A. (2003). A measure from line-of-sight magnetograms for prediction of coronal mass ejections. *Journal of Geophysical Research: Space Physics*, 108(A10). https://doi.org/10.1029/2003JA010030

Falconer, D. A., Moore, R. L., & Gary, G. A. (2008). Magnetogram Measures of Total Nonpotentiality for Prediction of Solar Coronal Mass Ejections from Active Regions of Any Degree of Magnetic Complexity. *The Astrophysical Journal*, 689(2), 1433–1442. https://doi.org/10.1086/591045

Florios, K., Kontogiannis, I., Park, S. H., Guerra, J. A., Benvenuto, F., Bloomfield, D. S., & Georgoulis, M. K. (2018). Forecasting Solar Flares Using Magnetogram-based Predictors and Machine Learning. *Solar Physics*, 293(2), 28. https://doi.org/10.1007/s11207-018-1250-4

Freiherr von Forstner, J. L., Guo, J., Wimmer-Schweingruber, R. F., Hassler, D. M., Temmer, M., Dumbović, M., … Zeitlin, C. J. (2018). Using Forbush Decreases to Derive the Transit Time of ICMEs Propagating from 1 AU to Mars. *Journal of Geophysical Research: Space Physics*, 123(1), 39–56. https://doi.org/10.1002/2017JA024700

Gopalswamy, N., Yashiro, A. S., Michalek, A. G., Stenborg, A. G., Vourlidas, A. A., Freeland, A. S., … Freeland, S. (2009). The SOHO/LASCO CME Catalog. *Earth Moon Planet*, 104, 295–313. https://doi.org/10.1007/s11038-008-9282-7

Gosling, J. T. (1993). The solar flare myth. *Journal of Geophysical Research: Space Physics*, 98(A11), 18937–18949. https://doi.org/10.1029/93ja01896

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. In *Neural Networks* (Vol. 18, pp. 602–610). Pergamon. https://doi.org/10.1016/j.neunet.2005.06.042

Hoeksema, J. T., Liu, Y., Hayashi, K., Sun, X., Schou, J., Couvidat, S., … Turmon, M. (2014). The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: Overview and Performance. *Solar Physics*, 289(9), 3483–3530. https://doi.org/10.1007/s11207-014-0516-8

Hudson, H., & Ryan, J. (1995). High-energy particles in solar flares. *Annual Review of Astronomy and Astrophysics*, 33(1), 239–282. https://doi.org/10.1146/annurev.aa.33.090195.001323

Inceoglu, F., Jeppesen, J. H., Kongstad, P., Marcano, N. J. H., Jacobsen, R. H., & Karoff, C. (2018). Using Machine Learning Methods to Forecast if Solar Flares Will Be Associated with CMEs and SEPs. *The Astrophysical Journal*, 861(2), 128. https://doi.org/10.3847/1538-4357/aac81e

Lang, K. R. (1995). Sun, Earth and Sky. Sun, Earth and Sky (1st ed.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-57852-6

Kumar, S., & Sharma, H. (2016). *A Survey on Decision Tree Algorithms of Classification in Data Mining*. *Article in International Journal of Science and Research* (Vol. 5). www.ijsr.net

Leka, K. D., & Barnes, G. (2003). Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. I. Data, General Approach, and Sample Results. *The Astrophysical Journal*, 595(2), 1277–1295. https://doi.org/10.1086/377511 (a)

Leka, K. D., & Barnes, G. (2003). Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. II. Discriminant Analysis. *The Astrophysical Journal*, 595(2), 1296–1306. https://doi.org/10.1086/377512 (b)

Liu, J., Ye, Y., Shen, C., Wang, Y., & Erdélyi, R. (2018). A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms: CAT-PUMA. *The Astrophysical Journal*, 855(2), 109. https://doi.org/10.3847/1538-4357/aaae69

Liu, H., Liu, C., Wang, J. T. L., & Wang, H. (2019). Predicting Solar Flares Using a Long Short-term Memory Network. *The Astrophysical Journal*, 877(2), 121. https://doi.org/10.3847/1538-4357/ab1b3c

Liu, H., Liu, C., Wang, J. T. L., & Wang, H. (2020). Predicting Coronal Mass Ejections Using SDO /HMI Vector Magnetic Data Products and Recurrent Neural Networks . *The Astrophysical Journal*, 890(1), 12. https://doi.org/10.3847/1538-4357/ab6850

Manoharan, P. K. (2006). Evolution of coronal mass ejections in the inner heliosphere: A study using white-light and scintillation images. *Solar Physics*, 235(1–2), 345–368. https://doi.org/10.1007/s11207-006-0100-y

Michalek, G., & Yashiro, S. (2013). CMEs and active regions on the sun. *Advances in Space Research*, 52(3), 521–527. https://doi.org/10.1016/j.asr.2013.04.001

Mumford, S. J., Christe, S., Pérez-Suárez, D., Ireland, J., Shih, A. Y., Inglis, A. R., … Kirk, M. S. (2015). SunPy - Python for solar physics. *Computational Science and Discovery*, 8(1), 014009. https://doi.org/10.1088/1749-4699/8/1/014009

Raheem, Abd-ur, Cavus, H., Coban, G.C., Kinaci, A. cumhur, 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. Monthly Notices of the Royal Astronomical Society, Volume 506, Issue 2, 1916–1926, https://doi.org/10.1093/mnras/stab1816

Raheem, A.-. ur ., Cavus, H., Coban, G. C., Kinaci, A. cumhur ., 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. doi:10.6084/m9.figshare.14512860.v4 [Dataset]

Richardson, I. G. (2014). Identification of Interplanetary Coronal Mass Ejections at Ulysses Using Multiple Solar Wind Signatures. *Solar Physics*, 289(10), 3843–3894. https://doi.org/10.1007/s11207-014-0540-8

Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, 19(1), 1–18. https://doi.org/10.1186/s12859-018-2451-4

Schrijver, C. J. (2007). A Characteristic Magnetic Field Pattern Associated with All Major Solar Flares and Its Use in Flare Forecasting. *The Astrophysical Journal*, 655(2), L117–L120. https://doi.org/10.1086/511857

Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1), 118–138. https://doi.org/10.1198/106186006X94072

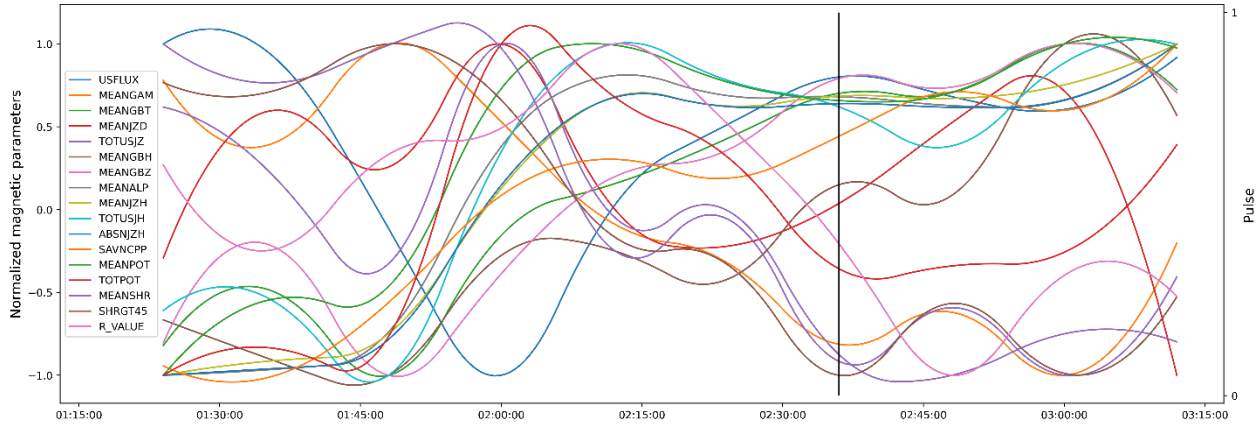
Tiwari, S. K., Venkatakrishnan, P., & Gosain, S. (2010). Magnetic non-potentiality of solar active regions and peak X-ray flux of the associated flares. *Astrophysical Journal*, 721(1), 622–629. https://doi.org/10.1088/0004-637X/721/1/622

Wang, J., Zhang, J., Deng, Y., Li, J., Tian, L., & Yang, X. (2002). Large-scale magnetic structures of coronal mass ejections. *Science in China, Series A: Mathematics, 45(SUPPL.)*, 57. https://doi.org/10.1007/bf02889685

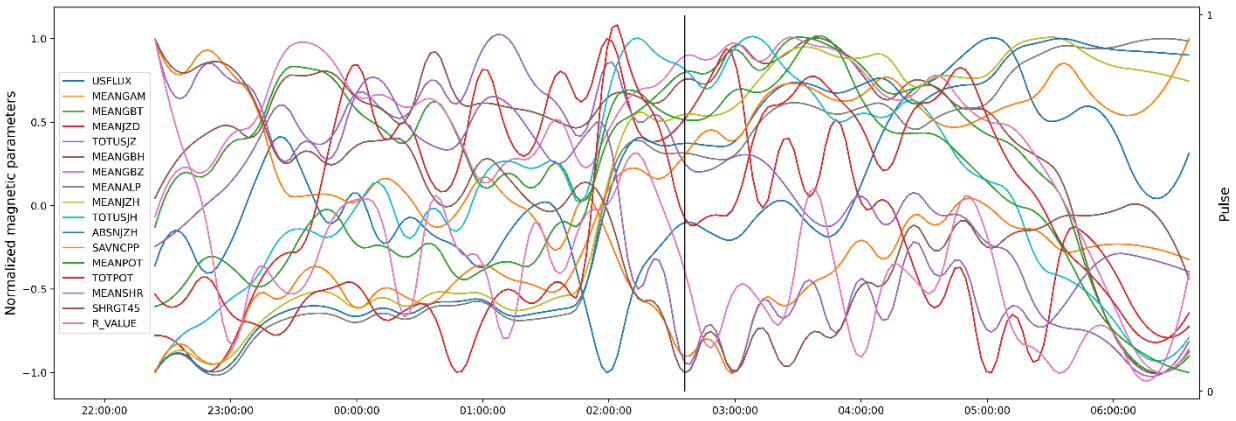
Yendapalli, V., Ruby, A. U., Theerthagiri, P., Jacob, I. J., & Vamsidhar, Y. (2020). Binary cross entropy with deep learning technique for Image classification A.Usha Ruby et al Binary cross entropy with deep learning technique for Image classification. *Article in International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 5393–5397. https://doi.org/10.30534/ijatcse/2020/175942020

# APPENDICES

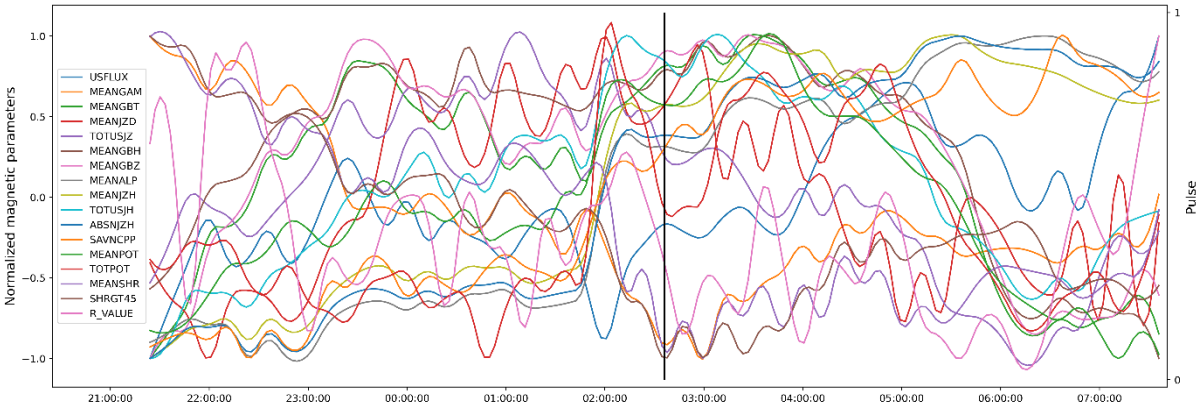
## APPENDIX 1. The representation of the 2-hour dataset as the first HARP is shown in here with its magnetic parameters along with the pulse generated for it based on CME timings from CDAW.




## APPENDIX 2. Same as appendix 1 but for 8-hour dataset.



## APPENDIX 3. Same as appendix 1 but for 10-hour dataset.




## APPENDIX 4. Same as appendix 1 but for 12-hour dataset.