ADVANCED ANALYTICAL TOOLS FOR GEOMAGNETIC STORM PREDICTION: ENSEMBLES AND THEIR INSIGHTS

by

TAYLOR K. LARKIN

DENISE J. MCMANUS, COMMITTEE CHAIR
BURCU B. KESKIN
VOLODYMYR MELNYKOV
CALI M. DAVIS
LE WANG

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Interdisciplinary Studies
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2017

Copyright Taylor K. Larkin 2017
ALL RIGHTS RESERVED

ABSTRACT

With the prevalence of technology found in modern society, the potential impact from strong geomagnetic storms cannot be underestimated. The primary drivers of these storms, coronal mass ejections (CMEs), are large explosions of solar material that are capable of being Earthward directed. Their effects have been well-studied in the literature, primarily by astrophysicists and others concerned with space weather. Because of the opportunity to obtain data about these phenomena, many studies have been done that build empirical models to predict if an impending CME will cause a strong geomagnetic disturbance. Two main types of data are typically used: data collected at the Sun and right before a CME's arrival to Earth. The former results in more timely predictions but less accuracy while the latter delivers better model performance but leaves much less time to prepare on Earth. Adequately dealing with this trade-off is still a growing area of research. In addition, creating and implementing advanced ensemble models for prediction and inference tasks based on machine and statistical learning theory have been lacking. Hence, this dissertation focuses on positing such models as well as present solutions for the trade-off problem.

The first work presents a new ensemble model based on random forests (RFs) to classify geoeffective CMEs using both types of data. This approach not only makes competitive predictions but also provides a reliable way to investigate the importance of the predictor variables (or features). The second work establishes a two-stage meta-learning framework that uses the first type of data in the first stage and both types in the second. The postulated method focuses on the trade-off problem, which is not addressed in the first work. The third work seeks to improve initial CME classification by incorporating CME image data. These images are introduced into a convolutional neural network (CNN) to create a set of deep learning features that increase the predictive power

of models that only use the first type of data. While the domain is specific to geomagnetic storm prediction, the methods proposed can be applied to a variety of prediction problems in other fields.

DEDICATION

I would like the dedicate this work to my family and friends, for without their support, this work could not have been done.

LIST OF ABBREVIATIONS AND SYMBOLS

ACE Advanced Composition Explorer.

AU astronomical unit.

AUC area under the receiver operating characteristic curve.

BDE bidirectional/counter-streaming solar wind suprathermal electron flow.

CIGRRF conditional inference guided regularized random forest.

CIRF conditional inference random forest.

CME coronal mass ejection.

CNN convolutional neural network.

DST Disturbance Storm Time index.

ESA European Space Agency.

GIC geomagnetically induced current.

GPU graphical processing unit.

GRF guided random forest.

GRRF guided regularized random forest.

ICME interplanetary coronal mass ejection.

IPI interplanetary information.

kappa kappa statistic.

km/s kilometers per second.

Kp Planetary K-index.

LASCO Large Angle and Spectrometric Coronagraph.

MC magnetic cloud.

NASA National Aeronautics and Space Administration.

NOAA National Oceanic and Atmospheric Administration.

nT nanoteslas.

OOB out-of-bag.

PGRF permutation guided random forest.

PRAUC area under the precision-recall curve.

PRF permutation random forest.

ReLU rectified linear unit.

RF random forest.

RMSE root mean square error.

RRF regularized random forest.

SCAD smoothly clipped absolute deviation.

SOHO Solar and Heliospheric Observatory.

STEREO Solar Terrestrial Relations Observatory.

SVM support vector machine.

WMAE weighted mean absolute error.

ACKNOWLEDGMENTS

I would like to thank God for blessing me with the opportunity and resources to attend school at the collegiate level. It is with His continued grace that I was able to study late into the night and better myself through the pursuit of knowledge. In addition, I would like to thank my family and friends for supporting me over the course of my years in school. Their positive impact on my education cannot be overstated.

I am also grateful for Drs. McManus, Keskin, Melnykov, Davis, and Wang for being a part of my dissertation committee and the Department of Information Systems, Statistics, and Management Science for serving as my home department. Of special note, I would like to recognize Dr. McManus, who served as the committee chair, for being my biggest advocate and inspirational source in the department. It is because of her that I sought and attained a Ph.D. Finally, I am appreciative of the support staff and other faculty in the department and in the graduate school for assisting me in completing my degree.

On a research note, I would like to thank NASA for their images and their creation of the CME catalog. This CME catalog is generated and maintained at the CDAW Data Center by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory. SOHO is a project of international cooperation between ESA and NASA. In addition, I would like to thank GSFC/SPDF, OMNIWeb, and NOAA for their public use databases and images. The majority of the data manipulation for this dissertation was done using SAS software. Copyright © 2013 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1

INTRODUCTION

Coronal mass ejections (CMEs) are colossal bursts of energy containing magnetic field and plasma components from the Sun. They can travel millions of miles per hour resulting in arrival times of just a few days if directed towards Earth. While generally these events do not cause significant harm, if they contain a strong magnetic field component opposite to that of Earth's, they can introduce harmful solar material into the atmosphere through a process called magnetic reconnection. The amassed energy in the upper atmosphere can cause severe issues on Earth such as worldwide black-outs, deteriorations of global positioning systems and radio communications, increased radiation levels, and physical damages to satellites. While only a small percentage of CMEs interact with Earth, it is not feasible for business entities to shut down operations for fear of experiencing these consequences. Hence, risk factor mitigation is an absolute necessity.

Much research has been conducted to try to assess if an impending CME will produce devastating effects or not. Typically, this is accomplished by collected data regarding a CME and creating empirical models to predict whether it will produce a geomagnetic storm (i.e., be geoeffective) and to what degree. In this way, the severity can be estimated prior to the CME reaching Earth. While both linear and non-linear approaches have been used, little work has been done to implement more advanced ensemble models. While these can provide superior performance in terms of accuracy, they are often difficult to interpret. Hence, it is important to offer insight as to which CME characteristics and solar wind parameters are most important for determining the geoeffectiveness of CMEs.

One of the major challenges with this approach is in regards to the type of data used. Because of interactions with the solar wind in the interplanetary medium, much of a CME's composition can be changed during its propagation towards Earth. Hence, using only the data recorded about a CME at its launch from the Sun can result in poor

predictions. On the other hand, using information collected much closer to Earth yields better forecasts; however, this is at the expense of lead time. Lead time is defined in this work as the time available for preparations to be made on Earth for an impending geomagnetic storm. The latter type of data results in only hours of lead time before to a CME's collision with Earth whereas the former gives at least a day. Finding a balance between these types of data remains of primary interest for geomagnetic storm prediction, which can be done using analytical methods.

In this dissertation, descriptive analytics is addressed through careful aggregation of several data sources so as to provide a more realistic prediction scenario in each work. Predictive analytics is applied through the creation of new models to both make future predictions and to gain insight as to why those predictions are made. Prescriptive analytics is implemented by suggesting a new two-stage framework that enables more timely predictions for geomagnetic storms.

The primary goal of the first work is to present an ensemble model that offers both predictive power and reliable insight as to which CME and solar wind predictor variables are most important. The posited approach, based on random forests (RFs), is referred to as a conditional inference guided regularized random forest (CIGRRF). CIGRRFs are a modification of guided regularized random forests (GRRFs), which build a forest of regularized random trees. The goal of such models is to regularize the information gain in a standard RF in order to provide a smaller set of predictor variables that are both relevant and non-redundant. However, because they rely on the Gini importance scheme to penalize each predictor variable, they are susceptible to biases when analyzing data of varying scale of measurement and number of categories, which is entirely possible when studying complex CME and solar wind information. Hence, a modification is proposed: regularize the information gain from using the permutation importance scheme from an unbiased random forest framework based on conditional inference. Using simulation studies, the CIGRRF shows less biased variable selection compared to the GRRF. Using a carefully constructed real CME dataset, the CIGRRF selects the sparsest solution while still maintaining predictive competitiveness with other classification techniques. The

main contributions here are the proposal of a less biased RF based variable selection method, creation of a comprehensive multiclass CME dataset with relevant information necessary for geomagnetic storm prediction, and the application of such an ensemble to assess the important drivers of CME-driven geomagnetic storms. Since the main focus of this work is to present an ensemble model with predictive and inferential capabilities, it does not address the time issue present with predicting geomagnetic storms.

Based on recent successes in using multi-step approaches, the primary goal of the second work is to address the time issues with a two-stage meta-learning framework. In the first stage, a CIGRRF is used to forecast an initial probability as to the geoffectiveness of a CME at its launch using data only collected near the Sun. Then, using these predicted probabilities, as well as both data collected near the Sun and near Earth, stacked generalization is used to make a final prediction quantifying the potential impact through the Disturbance Storm Time index (DST) value, which is a common metric used for measuring the strength of a geomagnetic storm. Stacked generalization is an ensemble strategy that incorporates the predictions from a diverse set of models (base-learners) by using them as inputs for another model (a meta-learner). The goal of this meta-learner is to deduce information about the biases from the base-learners and improve generalization to make more accurate predictions.

While the concept of stacked generalization is not new, this work posits the use of a regularized quantile regression model as the meta-learner in order to adequately deal with the presence of only a few strong CME events among many weak ones, which has been largely unexplored in the literature. In addition, in an effort to gain insight, a method is postulated for assessing variable importance in stacked generalization, an avenue that has received very little attention if at all. This involves calculating model-specific variable importance scores for each base-learner and then weighting these scores based on the coefficients from the meta-learner to produce a final aggregated variable importance score for each predictor variable. Using the real CME data, stacked generalization provides significantly better performance compared to using any one method alone. The main contributions here are the proposal of a two-stage approach for more timely geomag-

netic storm predictions, the creation of a larger CME dataset with relevant information where only a small number of observations produce strong geomagnetic disturbances, the application of regularized quantile regression as a meta-learner, and the proposal of a variable importance strategy for stacked generalization. While this framework does well in delivering more timely and accurate predictions, forecasting using only the stage one data remains a challenging task.

The primary goal of the third work is to improve upon the initial CME classifications made with data collected near the Sun using CME image information. While many works involve creating automated approaches for detecting CMEs from images of the Sun, very little work exists on making classifications regarding their geoeffectiveness. Of the studies that do exists, they require much preprocessing and generation of handcrafted features. Given the surge of deep learning in recent years, especially in regards to using convolutional neural networks (CNNs) for image classification tasks, a large opportunity exists to institute deep learning based approaches for CME classification. CNNs are deep neural networks that create higher-level abstractions of an image through use of convolutional layers. These layers can be connected to create a classification model or used as generic feature extractors.

Specifically in this work, a CNN is used as a feature extractor for a sample of CME images known to have interacted with Earth to generate new features to add alongside the traditional data used for CME classification. Nevertheless, because these features can be very high-dimensional, it is necessary to select only those that are informative for distinguishing strong CMEs from moderate or weak ones. The technique proposed in the first work could be used, however, it is not computationally tractable in high-dimensional settings. Hence, another RF based feature selection method is proposed using the idea of guided random forest (GRF). GRFs were originally created for circumventing the parallelization and correlation issues in GRRFs. While they are much faster than GRRFs, they still suffer from the bias issues related to the Gini importance scheme. Hence, similar to the first work, a modification is proposed, referred to as a permutation guided random forest (PGRF), to promote more unbiased feature selection using a permutation

importance framework designed for high-dimensional data. Two simulation and three real data studies are shown to demonstrate the improvement of PGRFs over GRFs in terms of speed and feature selection. Then, using real CME image data alongside their traditional characteristics, the performance of classifiers with and without using the PGRF selected deep learning features from a CNN is assessed. Results show that integrating these deep learning features offers better predictive performance. The main contributions here are the proposal of PGRFs, the creation of a CME data with relevant information including their images, the application of CNNs in geomagnetic storm prediction, and the idea of integrating PGRF selected deep learning features into initial CME classification.

2

A TREE ENSEMBLE MODEL FOR CLASSIFYING GEOEFFECTIVE CORONAL MASS EJECTIONS

2.1 Introduction

2.1.1 Coronal Mass Ejections

A CME is a massive explosion of magnetic field and plasma components from the Sun. Typically, CMEs travel at speeds between 400 and 1,000 kilometers per second (km/s) [48] resulting in an arrival time of around one to four days [82]; however, they can move as slowly as 100 km/s or as quickly as 3,000 km/s (or around 6.7 million miles per hour) [64]. These phenomena can contain a mass of solar material exceeding 10^{13} kilograms (or approximately 22 trillion pounds) [61] and can explode with the force of a billion hydrogen bombs [2]. Naturally, CME events are frequently associated with solar activity such as sunspots and solar flares [64]. During the solar minimum of the 11 year solar cycle (the period of time where the Sun has fewer sunspots and hence, weaker magnetic fields), CME events occur about once a day. During a solar maximum, this daily estimate increases to four or five per day. One plausible theory for why these incidents occur involves the Sun's need to release energy. As more sunspots develop, more coronal magnetic field structures become entangled; therefore, more energy is required to control the volatility and convolution. Once the energy surpasses a certain level, it becomes advantageous for the Sun to release these complex magnetic structures [48].

CMEs can be observed using coronagraphs, which obstruct the radiance of the Sun, much like in a total solar eclipse, so that the Sun's corona may be observed. The bright loops indicate the immense plasma clouds ejecting away from the Sun. Examples of some images from the Large Angle and Spectrometric Coronagraph (LASCO) can be seen in

Figure 2.1 courtesy of the National Aeronautics and Space Administration (NASA) and European Space Agency (ESA) Solar and Heliospheric Observatory (SOHO) mission [64]. With these instruments, NASA can categorize a CME as being a full or partial halo. Full halo CMEs appear to envelope the entire occulting disk of a coronagraph while partial halos only surround a portion of it. Halo CMEs are of special interest because they are more likely to reach and impact Earth [57].

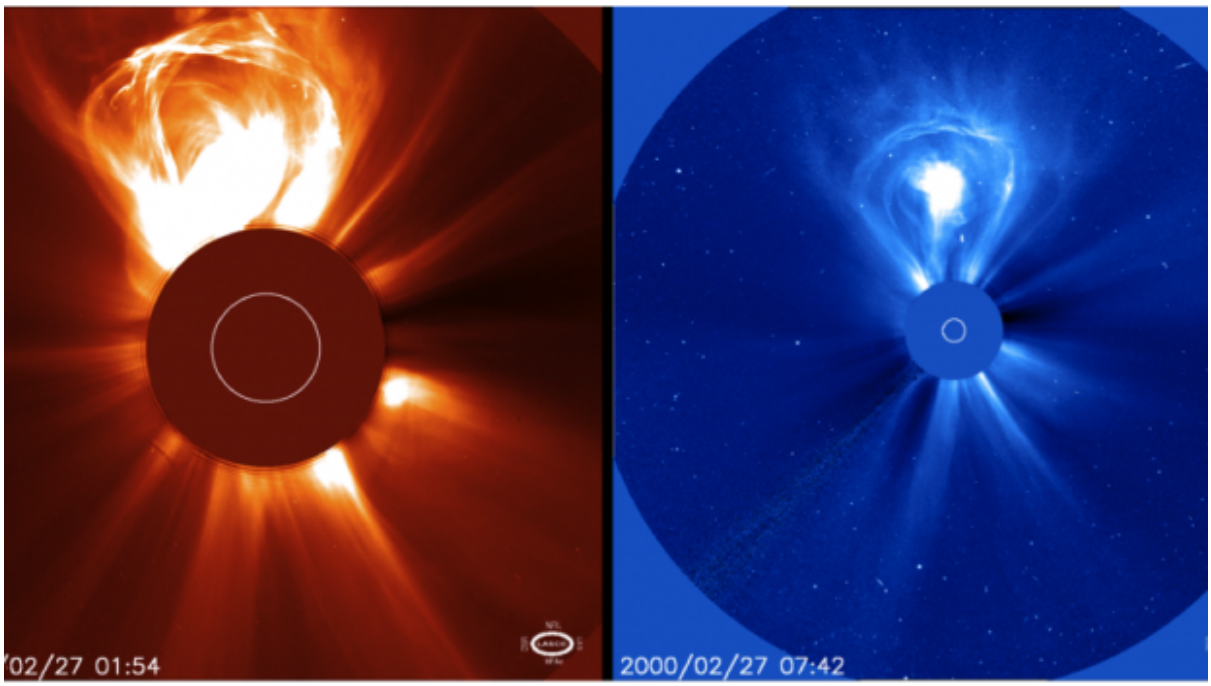


Figure 2.1: Coronagraph images of a partial halo CME from February 27, 2000.

As CMEs propagate outwardly through the heliosphere and interact with the neighboring solar wind, they become known as interplanetary coronal mass ejections (ICMEs). These interactions often change much of the composition of the original CME. Fortunately, details can be recorded *in-situ*, or nearby Earth, using spacecraft like the Advanced Composition Explorer (ACE) as the impending ICME passes over it. ICMEs are usually led by an initial shock and a sheath separated by heated and compressed plasma. Often, they are followed by magnetic clouds (MCs) [22] and bidirectional/counter-streaming solar wind suprathermal electron flows (BDEs) [42]. Exact definitions of when CMEs become ICMEs are unclear; regardless, a simplified distinction is that CMEs are observed at the Sun while ICMEs are detected at a greater distance away [48].

CMEs are often associated with solar flares [21] [91], which are sudden flashes of

light that catapult large amounts of energy into space. While these both involve massive bursts of energy from the Sun, they are different phenomena [1]. Solar flares are much faster as they travel through the interstellar medium at the speed of light and, in the worst-case scenario, can temporarily knockout radio communications. They are generally categorized into five major classes: A, B, C, M, and X, with X representing the most severe. NASA describes the difference between solar flares and CMEs with the following example [1]:

"The flare is like the muzzle flash, which can be seen anywhere in the vicinity. The CME is like the cannonball, propelled forward in a single, preferential direction... an immense cloud of magnetized particles hurled into space. Traveling over a million miles per hour, the hot material called plasma takes up to three days to reach Earth. The differences between the two types of explosions can be seen through solar telescopes, with flares appearing as a bright light and CMEs appearing as enormous fans of gas swelling into space."

When the force from a CME approaches Earth, it collides with the magnetosphere. This is the area encompassing Earth's magnetic field and serves as a line of defense against solar winds. The National Oceanic and Atmospheric Administration (NOAA) describes this event as "the appearance of water flowing around a rock in a stream" [65] as shown in Figure 2.2.

After the solar winds compress Earth's magnetic field on the day side (the side facing the Sun), they travel along the elongated magnetosphere into Earth's dark side (the side opposite of the Sun). The electrons are accelerated and energized in the tails of the magnetosphere. They filter down to the Polar Regions and clash with atmospheric gases causing geomagnetic storms. This energy transfer emits effulgences known as the *Aurora Borealis*, or Northern Lights, and the *Aurora Australis*, or Southern Lights, which can be seen near the poles [63].

In order to determine the strength of geomagnetic storms, a popular metric to use is the DST [86]. It is expressed in nanoteslas (nT) and recorded every hour from observatories around the world. Specifically, it measures the depression of the equatorial

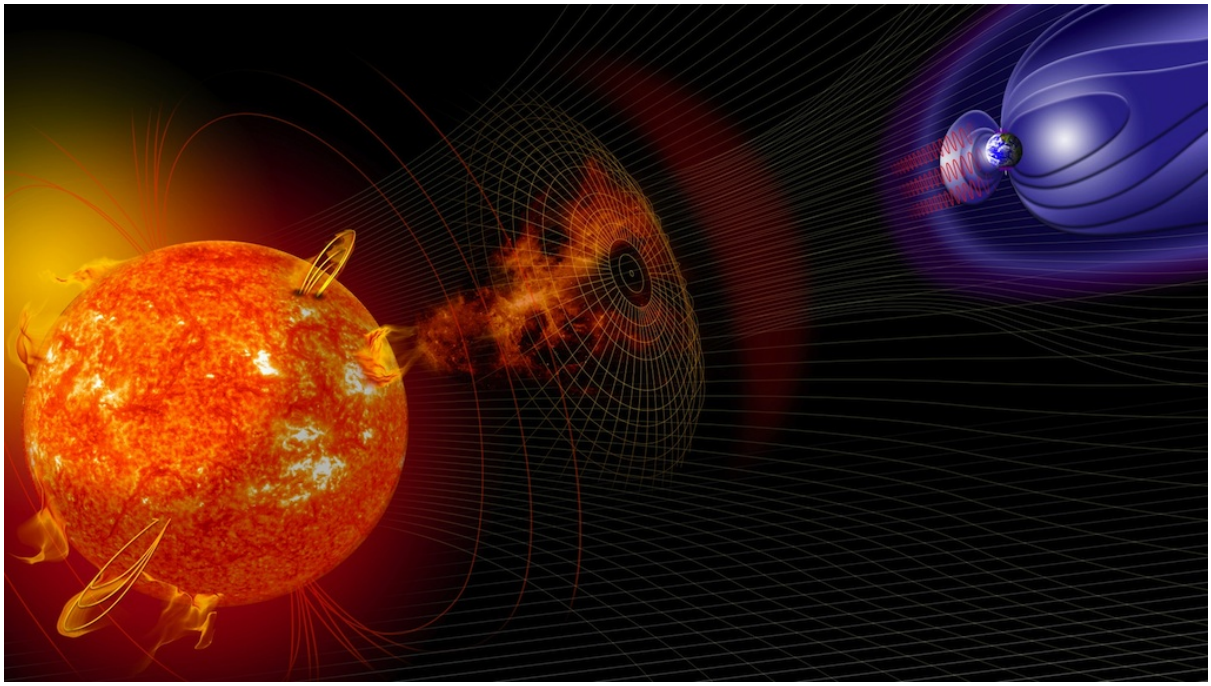


Figure 2.2: Rendering of Earth's magnetosphere interacting with solar wind from the Sun, courtesy of the NASA's Goddard Space Flight Center [3].

geomagnetic field, or horizontal component of the magnetic field; thus, the smaller the value of the DST, the more significant the disturbance on the magnetic field [48]. An alternative statistic is the Planetary K-index (Kp) [8]. Measured in integers from zero to nine, this provides a sense of the general variation in any component of the magnetic field through the use of an aggregated three hour interval estimate across each individual observatory. Higher values of Kp indicate stronger geomagnetic storms. NOAA has adopted this metric in their class definitions of geomagnetic storms.

2.1.2 Motivation for Studying CMEs

While responsible for the illustrious Northern Lights, geomagnetic storms have the potential to cause cataclysmic damage to Earth. Normally, the magnetic field is able to deflect most of the incoming plasma particles from the Sun. However, when an ICME contains a strong southward-directed magnetic field component, energy is transferred from the ICME 's magnetic field to Earth's through a process called magnetic reconnection [33] [35] [37] (as cited in [91]). Howard [48] describes two direct implications of this action:

1. "If the magnetic field of the ICME is directed southward (relative to the Earth), magnetic reconnection exposes the Earth to the plasma contained within the CME,

which is injected directly into the geomagnetic field. Reconnection causes closed field lines to open, accessing them to the solar wind and allowing a larger proportion of the Earth's atmosphere to be exposed to its plasma." (pg. 14)

2. "The increased pressure impacting the magnetosphere causes it to compress and closed magnetic field lines to be reduced in size. This results in a further expansion of the auroral ovals, where the effects of direct impact of solar wind particles with the atmosphere are exposed to more dense populations of people on Earth." (pg. 14)

Between increased plasma particles in Earth's geomagnetic field and a reduction of the magnetosphere towards the equator, more energy is amassed in the upper atmosphere, particularly at the poles. Consequently, this energy is impressed upon power transformers causing over-saturations and inducing black-outs via geomagnetically induced currents (GICs) [52]. Some other residuals of this over-accumulation of energy include the corrosion of pipelines, deteriorations of radio and global positioning system communications, radiation hazards in higher latitudes, orbiting satellite drag and damage to spacecraft software, and deficiencies in solar arrays [79]. These ramifications pose a significant threat to global telecommunications and electrical power infrastructures. From a economic perspective, risk factor mitigation is an absolute necessity within the global business environment [62].

2.1.3 Some Notable Geomagnetic Storm Events

On September 1, 1859, Richard Carrington and Richard Hodgson independently observed a solar flare outside of the city of London [26]. Carrington [24] described the event:

"My first impression was that by some chance a ray of light had penetrated a hole in the screen attached to the object-glass, by which the general image is thrown into shade, for the brilliancy was fully equal to that of direct sun-light." (pg. 13-14)

Little did they know that this solar flare was a precursor to a CME that produced the strongest geomagnetic disturbance on record. Noted as the "Carrington Event,"

this storm disrupted telegraph operations and communications worldwide [71] [11]. Its consequential aurora could be seen as far south as the Caribbean [69]. It appeared so bright that gold miners in the Rocky mountains began their day at 1 a.m., thinking it was morning time [69]. Odenwald, Green, and Taylor [68] estimate that if such an event were to occur in today's society, it would result in a financial impact amassing tens of billions of U.S. dollars (USD) due to damages toward commercial satellite systems. Moreover, it could leave 20-40 million U.S. citizens without power for extended periods of time with a estimated economic cost of $0.6-2.6 trillion USD [6].

On March 13, 1989, another geomagnetic storm affected North America and Northern Europe. Specifically, six million Québécois residents lost power for nine hours due to an over-saturation of transformers and equipment in the Hydro-Québec power system [4] (as cited in [12]). Such an event was estimated to cost more than $2 billion USD [9]. This crisis revitalized efforts for further exploration and research on how geomagnetically induced currents plague Earth's electrical systems [12].

The infamous "Halloween Storms," occurring during the last three days of October in 2003, caused minor interferences with power grids in North America. However, the same cannot be stated for the southern region of Sweden. Approximately 50,000 Malmö residents experienced loss of power due to GICs. This and other issues like public transit delays summed to a economic loss of about $0.5 million USD (or over four million Swedish krona) [72].

More recently in 2012, Earth narrowly avoided a potentially catastrophic geomagnetic storm. During the early morning of July 23, a CME catapulted from the Sun with an initial speed of 2,500 km/s. Satellites measured this extraordinarily fast CME *in-situ* some 19 hours later. While close enough to be measured by NASA satellites, it fortunately was not on track to strike earth. Baker et al. [7] explain that had this particular CME occurred a week prior, Earth would have been positioned in its direct path. Based on their assumptions, the ensuing geomagnetic storm would have been far more powerful and detrimental to modern society than the Carrington Event of 1859.

Although only a small fraction of these eruptions are directed towards Earth [48],

they are chiefly responsible for severe geomagnetic weather events [43] [13] [88] (as cited in [94]). Hence, it is imperative to study these phenomena and estimate their potential danger with innovative machine and statistical learning approaches.

2.1.4 The Purpose of This Work

CMEs can pose a significant threat upon the global business climate. While only a small percentage of CMEs approach Earth, it is not feasible for businesses to shut down power grids and telecommunication operations every time a CME is detected. Therefore, it is necessary to find accurate methods for predicting the geoeffectiveness, or the ability to impact Earth by producing a geomagnetic storm, of an impending CME. Additionally, it is advantageous to investigate the key indicators that determine this geoeffectiveness. Both of these tasks can be done via model based avenues such as with conditional inference random forests (CIRFs) [84], which are a variant of the popular RF [18] algorithm. Not only do these tree ensembles produce reliable variable importance measures, they enjoy the predictive advantages associated with RFs. These types of advanced models are largely absent in the literature for classifying CMEs. However, because CIRFs do not inherently reduce the feature space, opportunity exists to integrate these into variable selection frameworks. Therefore, the purpose of this work is to present a modified version of the CIRF that not only classifies the varying intensities of CMEs with competitive accuracy, but also delivers useful and parsimonious interpretation for practitioners on a real CME dataset.

The subsequent sections of this work read as follows. Section 2.2 briefly reviews some previous studies on predicting geoeffective CMEs as well as some additional background information beneficial for describing the methodology. Section 2.3 provides detailed explanations of how the dataset used is constructed, the proposed model, and the experimental strategy for both a simulation and real CME data study. Section 2.4 displays results of predictive performance and the resulting variable importance scores. Sections 2.5 and 2.6 conclude with a summary and postulates areas for future work.

2.2 Literature Review

2.2.1 Predicting Geoeffective CMEs

Many aspects of CMEs have been studied in the past for the purposes of prediction such as CME arrival time [41], when they will occur in the corona [73], and their effect on morbidity and mortality of humans [31]. The types of techniques implemented range from physics based models to empirical methods. Moreover, many have analyzed and summarized CME properties throughout the years [21] [91] [94] [82] [76]. While many works exist on predicting geomagnetic storms, few works exist using a binary classification systems [25], let alone extensions to true multiclass prediction frameworks. Furthermore, the inclusion of information from both CMEs and their interplanetary counterparts is necessary for geomagnetic space weather forecasting [32] [54]. Thus, emphasis is placed on reviewing those works focused on classification interpretations and incorporating data from both areas.

Srivastava [81] used logistic regression to study 64 geoeffective CMEs during the years of 1996-2002. As the response, she classified each event as either super-intense or not. By using a statistical model like logistic regression, inference on the importance of the predictor variables could be made. Her model correctly predicted approximately 78% (7 out of 9) CME events on a validation dataset and 85% on the respective training dataset. The results showed that the only statistically significant predictor variable in the analysis is the southward-directed magnetic field component of the ICME. This remains consistent with previous literature as the most prominent driver of geomagnetic storm strength. Unfortunately, only a small number of CMEs are able to be used for validation. A more comprehensive estimation of error could have been done via a resampling strategy [53].

Uwamahoro, McKinnell, and Habarulema [89] implemented a neural network on information from CME-driven geomagnetic storms between 1997 and 2006 to predict the probability of occurrence for geoeffective CMEs as a binary output. They divided their data into training (112 observations) and testing (43 observations) datasets comprised of partial and full halo CMEs. Their neural network identified 100% (19 out of 19)

of intense and 75% of moderate (18 out of 24) geomagnetic storms on the validation dataset. However, as indicated by the authors, their method only predicts the probability of a geomagnetic storm occurring, not the degree of geoeffectiveness the CMEs exhibit themselves.

Caswell and Rouleau [25] offered their own logistic regression model with backward-conditional model selection as well as training a series of artificial neural networks to predict the presence of daily geomagnetic storm activity for 577 days. Both models produced an accuracy rate of around 92%. Inference derived from the logistic regression model indicated that the interplanetary magnetic field component and solar wind plasma speed are the most significant predictors for storm occurrence. However, making inferences after model selection in this way is not completely valid [10].

Kim, Moon, Gopalswamy, Park, and Kim [54] instituted a two-step forecasting system for 55 CMEs from 1997-2003. In the first step, they predicted the strength of a geomagnetic storm using only CME information observed at the time of ejection. Then, the authors updated the forecast with the necessary interplanetary information to classify the intensity of the resulting storm from a given CME. The latter step was based on a proposed set of rules about the solar wind condition. This method contributed a medium-term to short-term forecast from the first observance of a CME to its approach to Earth. While this method yields accurate and interpretable results, the absence of using a validation scheme can lead to overfitting when predicting on future data [45].

2.2.2 Random Forests

Classification trees[1] [20] are a very popular technique for analyzing data due to their practical interpretations. They can identify useful splits on predictor variables in a dataset using binary recursive partitioning. In deciding these partitions, maximizing the information gain from the Gini index is typically used to determine which predictor variables to use. Some advantages to using classification trees are that they assume no distribution, produce interpretable logic rules, can handle varying scales and types of data as well as missing values, and are able to capture complex interactions [58]. However, tree models

[1]For simplicity, in this work only classification trees are discussed. However, it is important to note that the ideas presented can also apply to regression problems in the form of regression trees.

have high variance issues and are prone to overfitting [44]. To account for this weakness, Breiman [16] suggested it is advantageous to perform bagging, or bootstrap aggregation, in conjunction with tree models. That is, train many trees independently on bootstrap samples of the same size as the training data and aggregate the individual predictions by majority voting. To improve upon this procedure, Breiman [18] invoked randomness to posit the idea of RFs. RFs grow a plethora of trees that choose a random subset of predictor variables at each node to find the best split as opposed to using all of them as in a traditional classification trees or in the bagging procedure. The process is described in more detail below.

Define $\boldsymbol{X}$ to be the predictor matrix of dimension $n \times p$ and $\boldsymbol{Y}$ to be the vector of class labels of dimension $n \times 1$ for n observations and p predictor variables. Denote (x_i, y_i) as an observation for $i = 1, ..., n$ and x_j as a predictor variable for $j = 1, ..., p$. In particular,

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Furthermore, let *ntree* signify the number of trees used and *mtry* be the number of randomly selected predictor variables to choose from at each node v in constructing a RF. The algorithm is as follows [59]:

Step 1: Draw *ntree* random samples from the training data with replacement (bootstrap samples) of size n.

Step 2: For each of these *ntree* samples, fully grow a classification tree choosing among *mtry* at each v where $mtry \leq p$.

Step 3: Aggregate class predictions by majority voting across all trees.

The RF error rate depends on the correlation (how similar two trees are) and strength (how good of a classifier each tree is) of the forest [18] [19]. Better prediction results when less correlation and more strength exists amongst the trees. However, these situations are inverses of one another. That is, larger selections of *mtry* increase the strength of

each tree, but also boost the correlation and vice versa. Thus, it is advisable to test for an optimal *mtry* [19]. This methodology has many advantages including being able to run in parallel, manage thousands of variables without deletion, and be robust to noise and overfitting [19]. In addition, RFs can internally assess their error rate at the time of training using out-of-bag (OOB) samples. Specifically, during the construction of each tree, around a third of the observations are omitted from the bootstrap sample used to create a single tree. These withheld samples, the OOB samples, are then run down the respective tree. The classification accuracy is recorded and aggregated across all trees. In this way, each observation serves as an OOB sample approximately 36% of the time on average [59]. These OOB error estimates have been shown to be quite similar to those obtained from cross-validation [17] [44]. Therefore, in practice, a RF needs only to be trained once to gain a sense of performance on a validation dataset.

2.2.3 Variable Importance in Random Forests

RFs may also calculate variable importance scores for each predictor variable. This is usually done by using the aforementioned Gini index or using a permutation scheme. The former computes the mean information gain (or mean decrease in Gini node impurity) for each predictor variable in each split in all trees used. The latter randomly permutes values of a predictor variable from the OOB samples down the grown tree and measures the mean decrease in classification accuracy from the initial OOB sample accuracy over all trees. Define the Gini index at each v as

$$Gini(v) = \sum_{c=1}^{C} \mathcal{P}_c^v (1 - \mathcal{P}_c^v) \tag{2.1}$$

where C is the number of classes and $\mathcal{P}_c^v$ is the proportion of observations belonging to class c at node v. To calculate the information gain, the weighted average of the two descendant nodes v^L and v^R is subtracted from the parent node v. Thus, the more pure

the descendant nodes are, the more information gained at the parent node. Therefore,

$$\begin{aligned} Gain(x_j, v) = Gini(x_j, v) \\ - [\theta_L Gini(x_j, v^L) + \theta_R Gini(x_j, v^R)] \end{aligned} \tag{2.2}$$

where θ_L and θ_R represent the proportion of observations appointed to the left and right descendant nodes v^L and v^R, respectively, from the binary split. The x_j with the largest $Gain(x_j, v)$ is utilized for splitting. Define Imp_j as the importance score corresponding to predictor variable x_j. Thus,

$$\text{Imp}_j = \frac{\sum_{v \in S_{x_j}} Gain(x_j, v)}{ntree} \tag{2.3}$$

where S_{x_j} is the collection of nodes in which x_j is calculated to have the largest $Gain(x_j, v)$. That is, the total information gain by a particular predictor variable from each tree is summed across all trees and divided by the number of trees constructed.

In the permutation scheme, the motivation is that when randomly distributing the values of x_j (while holding the other predictor variables constant) in the OOB samples down the constructed tree, the original association between x_j and $\boldsymbol{Y}$ is broken [83]. Therefore, if x_j is truly influential on $\boldsymbol{Y}$, then the classification accuracy after the permutation should decrease compared to when $\boldsymbol{X}$ is non-permuted. Thus, this difference can be calculated for all x_j and used as a measure of variable importance [18]. Strobl, Boulesteix, Kneib, Augustin, and Zeileis [83] characterized this as so

$$\text{VI}^{(t)}(x_j) = \frac{\sum_{i \in \overline{\mathcal{B}}^{(t)}} I(y_i = \hat{y}_i^{(t)})}{|\overline{\mathcal{B}}^{(t)}|} - \frac{\sum_{i \in \overline{\mathcal{B}}^{(t)}} I(y_i = \hat{y}_i^{(t)'})}{|\overline{\mathcal{B}}^{(t)}|} \tag{2.4}$$

where

- $\overline{\mathcal{B}}^{(t)}$ represents the OOB sample for tree t such that $t \in \{1, ..., ntree\}$
- $\hat{y}_i^{(t)}$ $\{\hat{y}_i^{(t)'}\}$ is the predicted class label for the i^{th} observation before {after} permutation for tree t
- $|\overline{\mathcal{B}}^{(t)}|$ is the number of observations in the OOB sample for tree t

- $\sum_{i \in \overline{\mathcal{B}}^{(t)}} I(y_i = \hat{y}_i^{(t)})$ $\{\sum_{i \in \overline{\mathcal{B}}^{(t)}} I(y_i = \hat{y}_i^{(t)'})\}$ is the sum of correct classifications for the i observations in the OOB sample before {after} permutation for tree t

These can be aggregated in a similar manner as Eq. 2.3

$$\mathrm{VI}_j = \frac{\sum_{t=1}^{ntree} \mathrm{VI}^{(t)}(x_j)}{ntree} \tag{2.5}$$

While the permutation method yields more homogeneous importance scores across all predictor variables [44], more computation is necessary. Therefore, the Gini importance scheme can provide a fast approximation to variable importance that can be consistent to the permutation method [19].

2.2.4 Regularized Random Forests

An issue to take into consideration is the uniqueness of the predictor variables selected for splitting. Deng and Runger [28] expressed that classification trees may select redundant predictor variables since they are focused solely on optimizing the information gain. In other words, a tree may try to express the same piece of information by employing two predictor variables when really only one is necessary. This poses a problem when attempting to procure a set of unique predictor variables in situations where the goal is variable selection, such as in gene expression studies. Therefore, they proposed a regularized random forest (RRF). The concept of a RRF is to build a RF using regularized random trees [28]. These trees implement a penalty scheme so that new predictor variables are not used for splitting unless they add a considerable boost in information gain compared to those that have already been used for splitting. The authors defined a new information gain measure non-linearly as

$$Gain_R(x_j, v) = \begin{cases} \lambda \cdot Gain(x_j, v) & x_j \notin F \\ Gain(x_j, v) & x_j \in F \end{cases} \tag{2.6}$$

such that $\lambda \in (0, 1]$ where F indicates the collection of predictor variable selected in previous splits and λ, referred to as coefficient of regularization, is a penalty between zero to one. A RRF moves across each node in each tree and sequentially adds any x_j to

F if it can provide a substantial increase in information gain not already obtained. That is, $F = \emptyset$ before tree construction. Then, predictor variables are added to F when they deliver a higher information gain, while penalized, compared to those already in F until all the trees have been constructed. In essence,

1. $F = \{\} + x_1 \rightarrow F = \{x_1\} + x_4 \rightarrow F = \{x_1, x_4\} + x_3$

   next tree

2. $F = \{x_1, x_3, x_4\} \rightarrow F = \{x_1, x_3, x_4\} + x_5$

   $\vdots$

   $ntree$

This depicts a situation[2] where three unique predictor variables (x_1, x_3, x_4) are used for splitting in the first tree and predictor variable x_5 is added to F in the second tree. Note that in this situation, x_2 either is never randomly selected as x_j or it is unable to add enough non-redundant information to warrant its inclusion into F. As the λ term increases closer to one, a smaller penalty is applied to new predictor variables entering F and vice versa. Due to this regularization, it is expected that the predictor variables included in F are both relevant and non-redundant [28].

2.2.5 Guided Regularized Random Forests

In instances where the number of observations is small, the Gini information gain may have trouble determining the predictability of each predictor variable, since few observations may exist in the subsequent nodes. Deng and Runger [29] referred to this as the "node sparsity issue." In situations where the number of predictor variables is much larger than the number of observations, such as in gene selection problems [30] [95] [29] [27], many predictor variables may express the same information, even when maximum regularization is applied. When this occurs in a RRF, one predictor variable is randomly selected for the split. Thus, it is entirely possible to select a redundant variable.

To combat this point, Deng and Runger posited GRRFs [29]. The aim here is to exploit the variable importance scores from an initial run of a RF to help the RRF make

[2]Based on Figure 1 in [29]

more intelligent choices for splits as the trees are built sequentially, thereby, reducing the chance of having many predictor variables with the same predictive information at a node with few observations. Recall Eq. 2.6. In RRFs each predictor variable is assigned the same penalty λ for those not in F. However, for GRRFs, the regularized information gain becomes[3]

$$Gain_R(x_j, v) = \begin{cases} \lambda_j \cdot Gain(x_j, v) & x_j \notin F \\ Gain(x_j, v) & x_j \in F \end{cases} \tag{2.7}$$

where

$$\lambda_j = (1 - \gamma) + \gamma \hat{\text{Imp}}_j \quad \text{such that} \quad \gamma \in [0, 1] \tag{2.8}$$

and

$$\hat{\text{Imp}}_j = \frac{\text{Imp}_j}{max_{j=1}^{p}(\text{Imp}_j)} \quad \text{such that} \quad 0 \leq \hat{\text{Imp}}_j \leq 1 \tag{2.9}$$

Here, the formation of the penalties for each predictor variable is not uniform but adaptive. Those predictor variables with lower importance scores given by the initial RF are penalized more and vice versa. The parameter γ controls the impact of $\hat{\text{Imp}}_j$ on λ_j and is considered the importance coefficient. When $\gamma = 1$, the penalties are composed solely of the normalized variable importance scores. When $\gamma = 0$, the GRRF is equivalent to the least penalized RRF.

Experimental results on simulated and real gene datasets showed that GRRFs perform competitively with other variable selection methods such as lasso [87] and varSelRF [30], choose small subsets of predictor variables, and run with computational efficiency, especially compared to varSelRF which needs several RF executions. While both RRFs and GRRFs can be implemented as classifiers, Deng and Runger [29] noted that these models may have high variances compared to RFs because the trees are correlated. Hence, for prediction, they feed predictor variables selected by these procedures into another model,

[3]In the original work, Deng and Runger [29] initially propose $\lambda_j = (1-\gamma)\lambda_0 + \gamma \hat{\text{Imp}}_j$ where λ_0 controls the regularization intensity. Through their experiments, exploring the sensitivity of γ as opposed to λ_0 led to better predictive performance. Consequently, they set $\lambda_0 = 1$ for simplicity. Therefore, the same approach is adopted in this work.

like a RF, to estimate predictive performance.

2.2.6 Conditional Inference Random Forests

Although the Gini importance scheme is efficient, some biases exist. Strobl, Boulesteix, Zeileis, and Hothorn [84] showed via simulation studies that this selection process is biased in favor of continuous predictor variables and ones with many categories. Additionally, they find that the bootstrapping procedure further induces preferences towards predictor variables with more categories. As a result, they offered a solution by creating a forest of conditional inference trees [47], referred to as CIRFs, that use the permutation variable importance method.

As opposed to traditional classification trees, conditional inference trees split predictor variables not by maximizing the Gini information gain, but through the application of conditional inference independence tests, which are similar to χ^2 statistical tests. Following the unbiased recursive partitioning procedure proposed by Hothorn, Hornik, and Zeileis [47], reliable test statistics can be constructed, no matter whether the response and predictor variable being tested are of different forms (e.g. continuous versus binary or binary versus categorical). Comparing the p-values from the conditional distribution of these test statistics, the x_j with the smallest p-value out of $mtry$ is selected for splitting. In addition, the use of statistical tests helps the overfitting problem of classification trees by providing some stopping criteria (e.g. cease the recursive partitioning when the global null hypothesis of independence cannot be rejected for some pre-specified significance level). Together with subsampling, instead of bootstrapping, it is possible to achieve unbiased variable importance scores based on the permutation scheme [84]. That is, when the response and predictor variables are truly independent (or when the null hypothesis is true), each predictor variable has a equally likely chance of selection, despite their scale of measurement or number of categories [47]. Another advantage to these tree ensembles is that conditional variable importance [83] can be calculated to help increase the detection of influential predictor variables in the presence of multicollinearity, though this may be

computationally intensive for larger datasets.

2.2.7 Conditional Inference Guided Regularized Random Forests

While the CIRF implementation can provide better inferences for datasets with diverse scales of predictor variables, Archer and Kimes [5] demonstrated that if the data are standardized and continuous, the Gini importance scheme fairs well in distinguishing the important predictor variables. Considering many gene expression studies comprise data of this nature, the RRF and GRRF approaches are sensible for their intended purposes. However, if these methods are to be applied to data of varying scales of measurement and level, the bias issues associated with the Gini importance scheme need to be addressed. To improve upon this, this work posits the idea of CIGRRFs. That is, instead of using the normalized variable importance scores of a preliminary RF, substitute with those from a CIRF. This changes the penalized information gain in the following way:

$$Gain_R(x_j, v) = \begin{cases} \lambda_j \cdot Gain(x_j, v) & x_j \notin F \\ Gain(x_j, v) & x_j \in F \end{cases} \tag{2.10}$$

where

$$\lambda_j = (1 - \gamma) + \gamma \hat{\text{VI}}_j \quad \text{such that} \quad \gamma \in [0, 1] \tag{2.11}$$

and

$$\hat{\text{VI}}_j = \frac{\text{VI}_j - min_{j=1}^{P}(\text{VI}_j)}{max_{j=1}^{P}(\text{VI}_j) - min_{j=1}^{P}(\text{VI}_j)} \quad \text{such that} \quad 0 \leq \hat{\text{VI}}_j \leq 1 \tag{2.12}$$

Note that VI_j represents the permutation importance score from an initial execution of a CIRF for predictor variable x_j. Note further that these are min-max normalized. This is because permutation importance scores can sometimes be slightly negative due to randomness, so it is necessary to use this to keep them bounded between zero and one. By replacing the initial weights in a GRRF with those from a CIRF, a more reliable penalization scheme on those $x_j \notin F$ can be achieved as the trees are sequentially built. In turn, this will alleviate some of the bias issues imposed by the Gini information gain while still selecting relevant and non-redundant predictor variables. Given that CMEs and other space weather information can come in a variety of forms, it is necessary to

institute CIGRRFs instead of using RRFs or GRRFs. For predictive purposes, due to the high variance issues mentioned earlier, this work feeds predictor variables selected by the CIGRRF into a CIRF for classification since these have been shown to be unbiased, as opposed to using RFs as done by Deng and Runger [29]. The CIGRRF can be described in the following manner:

Step 1: Train a CIRF on the data.

Step 2: Calculate the unconditional[4] permutation variable importance score for each predictor variable from the CIRF built in step 1.

Step 3: Min-max normalize the importance scores from step 2.

Step 4: Train a GRRF using the importance scores from step 3 as the coefficients of regularization.

Step 5: Obtain the relevant and non-redundant predictor variables selected by the GRRF in step 4.

Furthermore, if a classifier can produce a similar error with less predictor variables, then it should be the preferred classifier [28]. Consequently, in this case, if a CIRF with CIGRRF variable selection can produce similar error compared to a CIRF using all the predictor variables, it makes sense to choose the more parsimonious approach. This can lead to better interpretations for those concerned with space weather, especially if a diverse set of predictor variables is considered.

2.3 Methodology

2.3.1 Data Preparation

ICME and CME properties are necessary for accurate predictions. The dataset used for experimentation is comprised of four sources of data: near-Earth ICME information provided by Richardson and Cane [23] [76]; OMNI hourly averaged solar wind data at

[4]This can be replaced with conditional variable importance scores [83] for possibly more adaptive penalties. However, because of the computational expense, only the unconditional variant is considered here.

one astronomical unit (AU) from the Coordinated Data Analysis (Workshop) Web [55]; CME measurements given by LASCO located on the SOHO satellite [39]; and some Sun phenomena recorded by NOAA [66].

2.3.2 Selecting CME Events

Given that only a small amount of CMEs approach Earth, it is important to study those that fall within its vicinity. Cane and Richardson [23] sought to compile a comprehensive list of ICMEs in close proximity to Earth between the years of 1996 and 2002. They updated this list to span through 2009 [76] and continue to report new ICMEs online. This dataset (referenced as "ICME list" in this work) has been utilized in many publications and is useful to the study of Earth-directed CMEs. A small subset of the data from 1997 with relevant information to this work can be found in Table 2.1. The dataset includes the detected arrival time of the disturbance from the ICME; time of the ICME's leading edge; time of the ICME's trailing edge; whether BDEs were present (Y = yes, N = no); evidence of MCs (2 = reported MC, 1 = evidence of rotation in magnetic field but cannot be fully defined as such, 0 = no presence of MCs, H = MC is from Huttunen, Schwenn, Bothmer, and Koskinen [50]); the minimum value in the DST that occurred within the ICME time interval; one AU transit speed from the associated CME to the disturbance time; and the associated CME from the LASCO catalog [39] (H = full halo CME, dg = data gap in LASCO catalog followed by the time of the assumed CME event).

Among the important predictor variables included, the most critical for this work is the identification of the most probable CME association with a respective ICME. This provides a justifiable connection based on expert opinion for combining the interplanetary measurements with the properties of a CME captured at the Sun. As noted by the authors, it can be difficult to determine exactly which CME is responsible for an ICME such as when multiple CMEs are detected at the around the same time. Therefore, only those observations where reasonable CME associations can be identified are considered

Disturbance MM/DD (UT)	Start MM/DD (UT)	End MM/DD (UT)	BDE	MC	DST (nt)	V_{tr} (km/S)	LASCO CME MM/DD (UT)
1997							
01/10 0104	01/10 0400	01/11 0200	Y	2	-78	507	01/06 1510 H
02/09 1321	02/10 0200	02/10 1900	Y	2	-68	683	02/07 0030 H
04/10 1745	04/11 0600	04/11 1900	Y	2	-82	552	04/07 1427 H
04/21 0600	04/21 1000	04/23 0400	Y	2	-107	...	
05/15 0159	05/15 0900	05/16 0000	N	2	-115	616	05/12 0530 H
05/26 0957	05/26 1600	05/27 1000	Y	2H	-74	381	05/21 2100
06/08 1636	06/08 1800	06/10 0000	Y	2	-84	...	
06/19 0032	06/19 0700	06/20 2300	Y	2	-36	...	
07/15 0311	07/15 0800	07/16 1100	Y	2	-45	...	
08/03 1042	08/03 1300	08/04 0300	Y	2	-48	410	07/30 0445 H
08/17 0200	08/17 0600	08/17 2000	N	0	-28	...	
09/03 0800	09/03 1300	09/03 2100	Y	1	-98	405	08/30 0130 H
09/21 1651	09/21 2100	09/22 1600	N	2	-36	450	09/17 2028 H
10/01 0059	10/01 1600	10/02 2300	Y	2	-98	580	09/28 0108 H
10/10 0300	10/10 1100	10/10 2200	Y	1	-64	...	
10/10 1612	10/10 2200	10/12 0000	Y	2	-130	430	10/06 1528
10/26 1200	10/27 0000	10/28 0700	Y	1	-60	572	10/23 1126 H
11/06 2248	11/07 0400	11/09 1200	Y	2	-110	640	11/04 0610 H
11/22 0949	11/22 1900	11/23 1400	Y	2	-108	640	dg (11/19 1700)
11/23 1900	11/24 0000	11/25 0000	Y	0	-47	...	
12/10 0526	12/10 1800	12/12 0000	...	0	-60	460	12/06 1027
12/30 0209	12/30 1000	12/31 1100	Y	1	-77	430	12/26 0231

Table 2.1: List of 2003 ICME events given in the catalog by Richardson and Cane [23] [76]. Full table can be found here: `http://www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm`.

(e.g., when the LASCO CME association is not missing).

2.3.3 Creation of the Multiclass Classification Problem

As indicated by Figure 2.3, the NOAA space weather scales range from G1 to G5 class storms and indicate minor to extreme cases, respectively. Kps less than four are labeled G0. Note the potential devastating effects for G5 class storms. Fortunately, these only occur on average four times per solar cycle, or four times every 11 years. However, as designated in the description, voltage corrections may be needed starting at G3 class storms. These occur around 200 times per solar cycle [67]. Loewe and Prölss [60] indicate through their study of more than 1,000 geomagnetic storms that the median maximum Kp value of strong storms (DST between -100nT and -200nT) is around seven and for weak storms (DST between -30nT and -50nT) is around four. Given this relationship, it is reasonable to infer that storms that achieve a minimum DST value of -100nT or lower can be considered a G3 class by the NOAA space weather scales and those larger than -50nT a G0 class. This relationship is important for the present study since many authors use the more granular DST to acquire interplanetary solar wind information.

Scale	Description	Effect	Physical measure
G 5	Extreme	**Power systems:** Widespread voltage control problems and protective system problems can occur, some grid systems may experience complete collapse or blackouts. Transformers may experience damage. **Spacecraft operations:** May experience extensive surface charging, problems with orientation, uplink/downlink and tracking satellites. **Other systems:** Pipeline currents can reach hundreds of amps, HF (high frequency) radio propagation may be impossible in many areas for one to two days, satellite navigation may be degraded for days, low-frequency radio navigation can be out for hours, and aurora has been seen as low as Florida and southern Texas (typically 40° geomagnetic lat.).	Kp = 9
G 4	Severe	**Power systems:** Possible widespread voltage control problems and some protective systems will mistakenly trip out key assets from the grid. **Spacecraft operations:** May experience surface charging and tracking problems, corrections may be needed for orientation problems. **Other systems:** Induced pipeline currents affect preventive measures, HF radio propagation sporadic, satellite navigation degraded for hours, low-frequency radio navigation disrupted, and aurora has been seen as low as Alabama and northern California (typically 45° geomagnetic lat.).	Kp = 8, including a 9-
G 3	Strong	**Power systems:** Voltage corrections may be required, false alarms triggered on some protection devices. **Spacecraft operations:** Surface charging may occur on satellite components, drag may increase on low-Earth-orbit satellites, and corrections may be needed for orientation problems. **Other systems:** Intermittent satellite navigation and low-frequency radio navigation problems may occur, HF radio may be intermittent, and aurora has been seen as low as Illinois and Oregon (typically 50° geomagnetic lat.).	Kp = 7
G 2	Moderate	**Power systems:** High-latitude power systems may experience voltage alarms, long-duration storms may cause transformer damage. **Spacecraft operations:** Corrective actions to orientation may be required by ground control; possible changes in drag affect orbit predictions. **Other systems:** HF radio propagation can fade at higher latitudes, and aurora has been seen as low as New York and Idaho (typically 55° geomagnetic lat.).	Kp = 6
G 1	Minor	**Power systems:** Weak power grid fluctuations can occur. **Spacecraft operations:** Minor impact on satellite operations possible. **Other systems:** Migratory animals are affected at this and higher levels; aurora is commonly visible at high latitudes (northern Michigan and Maine).	Kp = 5

Figure 2.3: NOAA space weather scales [67]. Not only are the scales assigned to a Kp value, but NOAA provides a short description of the effects of these storms at varying intensities.

Disturbance MM/DD (UT)	ICME Leading Edge MM/DD (UT)	ICME Trailing Edge MM/DD (UT)	LASCO CME MM/DD (UT)
1997			
01/10 0104	01/10 0400	01/11 0200	01/06 1510 H

Table 2.2: Sample of data from the ICME list. Starting from the disturbance time to the end of the trailing edge dictates the geomagnetic indices for the CME event.

Year	Month	Day	Hour	DST
...	...	...	...	...
1997	1	10	0	4
1997	1	10	1	18
1997	1	10	2	23
1997	1	10	3	13
1997	1	10	4	7
1997	1	10	5	-9
1997	1	10	6	-25
1997	1	10	7	-41
1997	1	10	8	-64
1997	***1***	***10***	***9***	***-78***
1997	1	10	10	-73
1997	1	10	11	-58
1997	1	10	12	-60
1997	1	10	13	-64
1997	1	10	14	-60
1997	1	10	15	-62
1997	1	10	16	-59
1997	1	10	17	-54
1997	1	10	18	-48
1997	1	10	19	-36
1997	1	10	20	-33
1997	1	10	21	-27
1997	1	10	22	-16
1997	1	10	23	-10
1997	1	11	0	3
1997	1	11	1	50
1997	1	11	2	46
...	...	...	...	...

Table 2.3: Sample of data from the OMNI dataset. Given the interval specified in Table 2.2, the lowest DST value is found and assigned to this CME event.

Year	Month	Day	Interval	Kp
...	...	...	...	...
1997	1	10	Kp Hours 00-03	2
1997	1	10	Kp Hours 03-06	4
1997	***1***	***10***	***Kp Hours 06-09***	***6***
1997	1	10	Kp Hours 09-12	6
1997	1	10	Kp Hours 12-15	4
1997	1	10	Kp Hours 15-18	3
1997	1	10	Kp Hours 18-21	3
1997	1	10	Kp Hours 21-24	2
1997	1	11	Kp Hours 00-03	4
...	...	...	...	...

Table 2.4: Sample of data from the NOAA dataset of Kp indices. Given the interval specified in Table 2.2, the highest Kp value is found and assigned to this CME event.

To assign a strength index to a CME event, a similar strategy utilized by Richardson and Cane [76] is implemented. An example of the process can be found via Tables 2.2, 2.3, and 2.4. The authors report the disturbance arrival time, the ICME magnetic field leading edge, and the respective trailing edge in their dataset. Therefore, starting at the disturbance time (or a little before) and ending at the trailing edge of the ICME (or a little after), the strength of an ICME can be realized by discerning the most severe geomagnetic index value from within that time frame (lowest DST or highest Kp). For comparison purposes, both values from different datasets are recorded for each selected ICME. Values of the DST and the Kp are provided by the OMNI database and NOAA [66], respectively. This process is conducted for each event with a probable CME association until all have been defined with a value for each index. While the ICME list provides the DST value, this is extracted along with the Kp for consistency. The DST value extracted matched quite well with those from the ICME list.

For time spans that overlap, the DST value documented by Richardson and Cane assisted in distinguishing values. That is, when the interval of one ICME event protruded into another (thereby making identification partially ambiguous), the ICME with the lower DST is given the higher Kp. The other would receive the second highest Kp within its time interval. If the minimum DST from the OMNI dataset falls on the arrival time specified by the authors, default is given to the authors' recorded DST value in the ICME list. If it is clear that a DST minimum realizes right before the arrival time, that value is recorded. This occurrence is entirely possible since some of the reported disturbance times are listed at the same time as the approximated leading edge. Once the DST and Kp indices are established, the multiclass classification problem can be formulated with the help of the NOAA geomagnetic storm definitions. For comparison purposes, each ICME is given a classification according to the DST and another given by Kp. That is, those with a DST of -100nT or less are labeled strongly geoeffective by the DST and those with a Kp of seven or higher are labeled likewise. Those with a DST of -50nT or more and Kp of four or less are labeled weakly geoeffective. Otherwise, it is considered moderately geoeffective.

2.3.4 ICME Predictor Variables

In an effort to incorporate the most important interplanetary parameters for the analysis, predictor variables from the OMNI database discussed by Kim et al. [54] are chosen. They include the southward magnetic field component (B_z), interplanetary electric field (E_y), plasma flow speed (V_{sw}), proton temperature (T_p), proton density (D_p), and flow pressure (P). To select these values, a similar procedure described by the authors is adopted here. Beginning at the disturbance time from Richardson and Cane [76] and ending through one hour before the DST minimum, the lowest B_z value is recorded. For simplicity, all of the other solar wind parameter values are taken at the hour of the B_z minimum. An example of this can be found via Table 2.5. Utilizing values from at least one hour before the DST minimum gives a lead time prior to the climax of the geomagnetic storms and allows for a more realistic prediction scenario. In addition to these from the OMNI database, three predictor variables from the ICME list are included: whether or not BDEs are observed, whether or not a MC is reported, and the shock transit speed to one AU.

2.3.5 CME and Solar Predictor Variables

To incorporate initial CME properties, data from the LASCO/SOHO catalog [39] is joined on the probable CME association. For those CME associations in which time discrepancies exist between the LASCO/SOHO catalog and the ICME list, the closest CME is chosen. For instance, according to the LASCO instrument, at 21:30:08 UT on February 17th, 2000, a full halo CME was recorded. In the ICME list, the authors indicated a full halo CME association at 20:06:00 UT on that same day. Given that no other full halo CME occurred on that day, it stands to reason that the official time of this CME in the LASCO/SOHO catalog may have changed since the authors collected the information. In addition, for circumstances in which LASCO detected multiple CMEs at the exact same time, the most likely CME is chosen based on the remarks in the LASCO/SOHO catalog. For example, one may be commented as a "poor event," so consequently, the other is elected as the probable CME association (see August 6th, 2000 at 18:30:32 UT). Due to these minor issues, the shock transit speed to one AU in

Year	Month	Day	Hour	DST	B_z	E_y	V_{sw}	T_p	D_p	P
...	...	...	...	...	...	...	...	...	...	...
1997	1	10	0	4	0	-0.07	375	43252	7.2	1.76
1997	1	10	1	18	3.4	-1.7	415	112809	12.4	3.92
1997	1	10	2	23	-3.3	0.92	437	112811	11.8	4.23
1997	1	10	3	13	-0.5	-0.27	442	171246	13.7	5.31
1997	1	10	4	7	-2.9	0.83	463	191691	11.9	4.74
1997	1	10	5	-9	-11.3	4.86	467	43343	3.4	1.29
1997	1	10	6	-25	-13.4	6.04	465	48410	3.5	1.32
1997	1	10	7	-41	***-15.3***	***7.02***	***459***	***48077***	***7.8***	***2.84***
1997	1	10	8	-64	-15.1	6.84	453	51249	7.3	2.61
1997	***1***	***10***	***9***	***-78***	-13.1	5.74	445	40751	4.7	1.64
1997	1	10	10	-73	-11	4.71	444	40093	6.7	2.36
1997	1	10	11	-58	-10	4.23	441	44444	4.9	1.67
1997	1	10	12	-60	-8.9	3.94	453	53045	5.1	1.89
1997	1	10	13	-64	-6.8	3.07	452	60124	5.3	1.96
1997	1	10	14	-60	-4.8	2.36	454	58219	6.4	2.35
1997	1	10	15	-62	-3.6	2	454	58159	5.7	2.11
1997	1	10	16	-59	-1.9	1.47	444	75908	6.6	2.33
1997	1	10	17	-54	0.8	0.53	443	64900	7.3	2.56
1997	1	10	18	-48	1.1	0.73	430	48016	6.4	2.09
1997	1	10	19	-36	1.7	0.8	420	19647	9.8	2.99
1997	1	10	20	-33	3.3	0.33	418	19788	13.6	4.13
1997	1	10	21	-27	10.4	-2.95	416	26943	18	5.45
1997	1	10	22	-16	10	-2.64	413	25068	20.9	6.24
1997	1	10	23	-10	9	-2.18	411	24854	26.7	7.89
1997	1	11	0	3	10.4	-2.94	414	27662	38.1	12
1997	1	11	1	50	14.8	-5.28	416	25890	54.7	19.54
1997	1	11	2	46	18.2	-7.06	413	31800	61.2	19.18
...	...	...	...	...	...	...	...	...	...	...

Table 2.5: Sample of data from the OMNI dataset displaying the selection procedure for the ICME predictor variables. The minimum B_z value usually occurs before the minimum DST value as shown here.

the ICME list is recalculated. Gopalswamy et al. [39] viewed the rudimentary elements of CMEs as their linear speed, angular width, central position angle, and acceleration. Because central position angle does not exist for full halo CMEs, measurement position angle is used instead.

Additionally, daily solar weather information is obtained from NOAA [66]. These include the average solar radio flux, number of sunspots, sunspot area, number of new sunspot regions, and a count of the C, M, and X-class solar flares for that day. Since these data are daily aggregations, they are merged with the CME data by the day in which the CME was expelled from the Sun. Finally, in order to preserve the validity of the data, any observation with missing data is omitted from the compiled final dataset (see Tables 2.6 and 2.7 for a sample of this data). The resulting efforts in the previous sections yield a dataset containing 179 CMEs with 20 predictor variables characterizing their properties at different stages of life spanning from January of 1997 through October of 2013. The list of predictor variables can be found in Table 2.8. Note that these vary in type, hence, illustrating the necessity to use the proposed CIGRRF for variable selection.

2.3.6 Data Exploration

Of the 179 chosen CME events, 59 are defined as producing strong geomagnetic storms using the DST and 53 using the Kp criterion. The DST and Kp classifications agree on about 73% (130) of all events in the dataset. Of the 49 storms not in agreement, 30 of them occur for ICME events that are classified as producing a moderate geomagnetic storm by the Kp but determined as either strong or weak according to the DST. Notably, this difference is most likely due to the fact that the maximum Kp usually precedes the DST minimum by one to two hours [60]. Therefore, those events on the borderline of being classified as strong or weak storms result in the largest discrepancy. This can be seen in Figure 2.4 as the largest disagreements occur in the off-diagonal entries of the moderate row and column. Regardless, the distributions listed in Table 2.9 are similar. However, because the distribution across each category is more even, the DST classifications will

LASCO CME MM/DD/YYYY UT	Disturbance Arrival MM/DD/YYYY UT	ICME Leading Edge MM/DD UT	ICME Trailing Edge MM/DD UT	Kp Index	Geoeffectiveness (Kp)	DST	Geoeffectiveness (DST)
01/06/1997 15:10:42	01/10/1997 01:04	01/10 04:00	01/11 02:00	6	moderate	-78	moderate
02/07/1997 00:30:05	02/09/1997 13:21	02/10 02:00	02/10 19:00	5	moderate	-68	moderate
04/07/1997 14:27:44	04/10/1997 17:45	04/11 06:00	04/11 19:00	7	strong	-82	moderate
05/12/1997 05:30:05	05/15/1997 01:59	05/15 09:00	05/16 00:00	7	strong	-115	strong
05/21/1997 21:00:53	05/26/1997 09:57	05/26 16:00	05/27 10:00	6	moderate	-73	moderate
07/30/1997 04:45:47	08/03/1997 10:42	08/03 13:00	08/04 03:00	5	moderate	-49	weak
08/30/1997 01:30:35	09/03/1997 08:00	09/03 13:00	09/03 21:00	5	moderate	-98	moderate
09/17/1997 20:28:48	09/21/1997 16:51	09/21 21:00	09/22 16:00	4	weak	-36	weak
09/28/1997 01:08:33	10/01/1997 00:59	10/01 16:00	10/02 23:00	6	moderate	-98	moderate
10/06/1997 15:28:20	10/10/1997 16:12	10/10 22:00	10/12 00:00	6	moderate	-130	strong
10/23/1997 11:26:50	10/26/1997 12:00	10/27 00:00	10/28 07:00	4	weak	-60	moderate
11/04/1997 06:10:05	11/06/1997 22:48	11/07 04:00	11/09 12:00	7	strong	-110	strong
12/26/1997 02:31:54	12/30/1997 02:09	12/30 10:00	12/31 11:00	6	moderate	-77	moderate
01/02/1998 23:28:20	01/06/1998 14:16	01/07 01:00	01/08 22:00	6	moderate	-77	moderate
01/17/1998 04:09:20	01/21/1998 04:00	01/21 06:00	01/22 13:00	4	weak	-11	weak
01/25/1998 15:26:34	01/28/1998 16:00	01/29 20:00	01/31 01:00	5	moderate	-55	moderate
02/14/1998 06:55:05	02/17/1998 04:00	02/17 10:00	02/17 21:00	5	moderate	-100	strong
02/28/1998 12:48:00	03/04/1998 11:56	03/04 13:00	03/06 09:00	3	weak	-36	weak
04/29/1998 16:58:54	05/01/1998 21:56	05/02 05:00	05/04 02:00	7	strong	-85	moderate
05/02/1998 14:06:12	05/04/1998 02:15	05/04 10:00	05/07 23:00	9	strong	-205	strong
06/21/1998 05:35:10	06/24/1998 10:00	06/24 16:00	06/25 23:00	4	weak	-25	weak
10/15/1998 10:04:36	10/18/1998 19:52	10/19 04:00	10/20 07:00	6	moderate	-112	strong
11/04/1998 07:54:06	11/07/1998 08:15	11/07 22:00	11/09 01:00	7	strong	-81	moderate
11/05/1998 20:44:02	11/08/1998 04:51	11/09 01:00	11/11 01:00	8	strong	-149	strong
11/09/1998 18:17:55	11/13/1998 01:43	11/13 02:00	11/14 12:00	6	moderate	-131	strong
04/13/1999 03:30:05	04/16/1999 11:25	04/16 18:00	04/17 19:00	7	strong	-91	moderate
04/18/1999 08:30:05	04/20/1999 16:00	04/21 04:00	04/22 14:00	4	weak	-31	weak
06/24/1999 13:31:24	06/26/1999 20:16	06/27 22:00	06/29 04:00	6	moderate	-41	weak
07/03/1999 19:54:05	07/06/1999 15:09	07/06 21:00	07/07 02:00	3	weak	-1	weak
07/23/1999 21:30:09	07/26/1999 23:33	07/27 17:00	07/29 12:00	3	weak	-38	weak
07/28/1999 04:30:05	07/30/1999 16:00	07/30 20:00	07/31 08:00	6	moderate	-53	moderate
07/28/1999 09:06:05	07/31/1999 18:37	07/31 19:00	08/02 06:00	4	weak	-39	weak
08/09/1999 03:26:05	08/11/1999 23:00	08/12 03:00	08/14 00:00	4	weak	-13	weak
08/17/1999 13:31:51	08/20/1999 23:00	08/20 23:00	08/23 11:00	6	moderate	-66	moderate
09/20/1999 06:06:05	09/22/1999 12:22	09/22 19:00	09/24 03:00	7	strong	-173	strong
10/18/1999 00:06:06	10/21/1999 02:25	10/21 08:00	10/22 07:00	8	strong	-237	strong

Table 2.6: A sample of the first eight columns in the compiled dataset for CMEs occurring between 1997 and 1999.

LASCO CME MM/DD/YYYY UT	B_z	E_y	V_{sw}	T_p	D_p	P	BDE	MC	TV	AW	LS	Acc	MPA	$RFlux$	SSN	SSA	NR	$XrayC$	$XrayM$	$XrayX$
01/06/1997 15:10:42	-15.30	7.02	459	48077	7.8	2.84	1	1	507	360	136	4.1	180	73	15	10	0	0	0	0
02/07/1997 00:30:05	-7.20	3.31	479	72722	0.3	0.13	1	1	683	360	490	14.3	266	76	38	30	0	0	0	0
04/07/1997 14:27:44	-7.60	3.82	444	225286	17.6	6.67	1	1	552	360	878	3.3	123	77	16	30	0	1	0	0
05/12/1997 05:30:05	-24.60	10.36	428	53612	5.8	2.21	0	1	607	360	464	15	264	72	12	60	0	1	0	0
05/21/1997 21:00:53	-10.60	3.23	336	45817	7.5	1.48	1	1	381	165	296	1.4	267	85	79	260	1	1	1	0
07/30/1997 04:45:47	-10.90	6.30	485	32798	4.6	1.89	1	1	408	360	104	0.8	269	71	12	0	0	0	0	0
08/30/1997 01:30:35	-10.90	5.03	412	48063	7.6	2.8	1	0	405	360	371	9.3	67	92	69	220	1	1	0	0
09/17/1997 20:28:48	-3.70	2.48	459	69244	16.9	6.28	0	1	450	360	377	0	263	93	52	490	0	6	2	0
09/28/1997 01:08:33	-7.40	4.61	490	155500	6.2	2.75	1	1	578	360	359	2.8	87	87	23	260	0	1	0	0
10/06/1997 15:28:20	-10.40	4.60	438	55510	21.3	8.03	1	1	430	174	293	15.9	130	84	24	20	0	0	0	0
10/23/1997 11:26:50	-6.80	2.59	462	32321	3.4	1.55	1	0	573	360	503	3.7	305	80	0	0	0	0	0	0
11/04/1997 06:10:05	-12.20	5.78	448	375732	19.5	7.83	1	1	643	360	785	22.1	243	118	68	1030	1	13	2	1
12/26/1997 02:31:54	-10.70	3.64	350	35598	10.1	2.13	1	0	435	230	197	5.5	106	105	50	460	1	1	0	0
01/02/1998 23:28:20	-12.60	5.02	395	52209	18.4	4.91	1	1	479	360	438	6.5	275	101	50	290	0	2	1	0
01/17/1998 04:09:20	-4.10	1.99	452	32063	32.4	11.28	0	0	434	360	350	5.6	82	96	89	330	1	0	0	0
01/25/1998 15:26:34	-7.30	2.95	398	31856	5.6	1.6	1	0	573	360	693	7.4	112	108	104	350	1	2	1	0
02/14/1998 06:55:05	-12.70	3.88	400	23250	8.2	2.29	1	1	602	206	123	0.7	164	105	88	390	1	0	0	0
02/28/1998 12:48:00	-6.90	1.51	387	11969	13.6	3.51	0	1	437	169	176	2.9	282	94	60	100	0	0	0	0
04/29/1998 16:58:54	-12.50	7.28	617	24685	11.8	11.13	1	1	785	360	1374	44.8	336	101	65	250	1	2	1	0
05/02/1998 14:06:12	-19.90	24.16	833	659634	5.4	8.36	1	0	1150	360	938	28.8	331	117	110	660	1	6	0	1
06/21/1998 05:35:10	-4.90	3.15	534	83350	9.8	4.89	1	1	544	78	148	3.2	286	102	87	350	0	5	0	0
10/15/1998 10:04:36	-22.70	6.53	416	14527	3	0.98	1	1	508	360	262	3.2	264	131	113	450	3	2	0	0
11/04/1998 07:54:06	-4.80	4.06	514	94682	10.8	5.49	1	0	574	360	523	19.6	349	141	124	620	2	11	0	0
11/05/1998 20:44:02	-11.40	5.91	462	7359	5.5	2.04	1	1	741	360	1118	24	300	153	137	680	1	11	3	0
11/09/1998 18:17:55	-17.50	6.41	377	30002	9.3	2.31	1	1	523	190	325	2.6	338	162	107	1310	1	7	0	0
04/13/1999 03:30:05	-15.20	4.88	428	24685	14.3	6.44	1	1	520	261	291	0.2	194	130	122	210	1	1	0	0
04/18/1999 08:30:05	-5.30	1.89	511	223005	8.1	4.07	1	1	749	112	475	12.4	53	113	98	220	0	0	0	0
06/24/1999 13:31:24	-4.90	3.32	511	313810	4.3	2.33	1	0	759	360	975	32.4	335	185	229	1530	3	13	0	0
07/03/1999 19:54:05	5.10	-2.29	457	300430	5.1	2.12	0	0	618	139	536	3.1	303	197	213	1400	3	8	0	0
07/23/1999 21:30:09	-5.70	2.30	389	9674	5.2	1.48	1	0	561	123	329	0.9	52	194	158	1210	1	5	3	0
07/28/1999 04:30:05	-10.30	6.86	641	60747	44.8	41.68	1	0	698	189	361	0.8	96	198	218	1540	2	4	2	0
07/28/1999 09:06:05	-6.10	3.78	619	45031	1.7	1.31	0	0	510	360	462	1.9	6	198	218	1540	2	4	2	0
08/09/1999 03:26:05	-3.30	2.07	406	47628	8.6	2.55	1	0	615	212	395	4.1	183	138	118	550	0	6	0	0
08/17/1999 13:31:51	-7.80	3.16	416	5590	2.6	0.81	1	0	510	261	776	45.7	46	141	67	460	0	3	0	0
09/20/1999 06:06:05	-18.50	8.84	589	109689	24.3	17.4	1	0	766	360	604	14.5	14	145	78	350	0	7	0	0
10/18/1999 00:06:06	-28.20	16.24	529	45018	8.6	4.39	1	0	559	240	144	3.5	184	173	135	1360	0	8	0	0

Table 2.7: All of the other columns included in compiled dataset. These will serve as the final set of predictor variables for classifying geoeffective CMEs (see Table 2.8). All predictor variables for modeling are quantitative except BDE and MC which are converted into binary variables. The first column from Table 2.6 is repeated for readability.

x	Type	Description
B_z	C	Southward magnetic field component in nanoteslas (nT)
E_y	C	Interplanetary electric field in millivolts per meter (mV/m)
V_{sw}	C	Plasma flow speed in kilometers per second (km/s)
T_p	C	Proton temperature in degrees Kelvin (K)
D_p	C	Proton density in Newtons per cubic centimeter (N/cm^3)
P	C	Flow pressure in nanopascals (nPa)
BDE	B	Evidence of BDEs
MC	B	Reported association of MC structure
TV	C	1 AU transit speed from CME to geomagnetic disturbance (km/s)
AW	C	Sky-plane width of CME in degrees
LS	C	Linear speed of CME (km/s)
Acc	C	Acceleration of CME in meters per seconds squared (m/s^2)
MPA	C	Measurement position angle of CME at the height-time measurements in degrees
$RFlux$	C	Daily average 10.7cm flux values of solar radio emissions on CME ejection day in $10^{-22}J/s/m^2/Hz$
SSN	D	Number of sunspots recorded on CME ejection day
SSA	C	Sum of the corrected area of all observed sunspots on CME ejection day in millionths of the solar hemisphere
NR	D	Number of new sunspot regions on CME ejection day
$XrayC$	D	Number of C-class solar flares on CME ejection day
$XrayM$	D	Number of M-class solar flares on CME ejection day
$XrayX$	D	Number of X-class solar flares on CME ejection day

Table 2.8: List of predictor variables for modeling. Types are indicated as C-Continuous, D-Discrete, B-Binary.

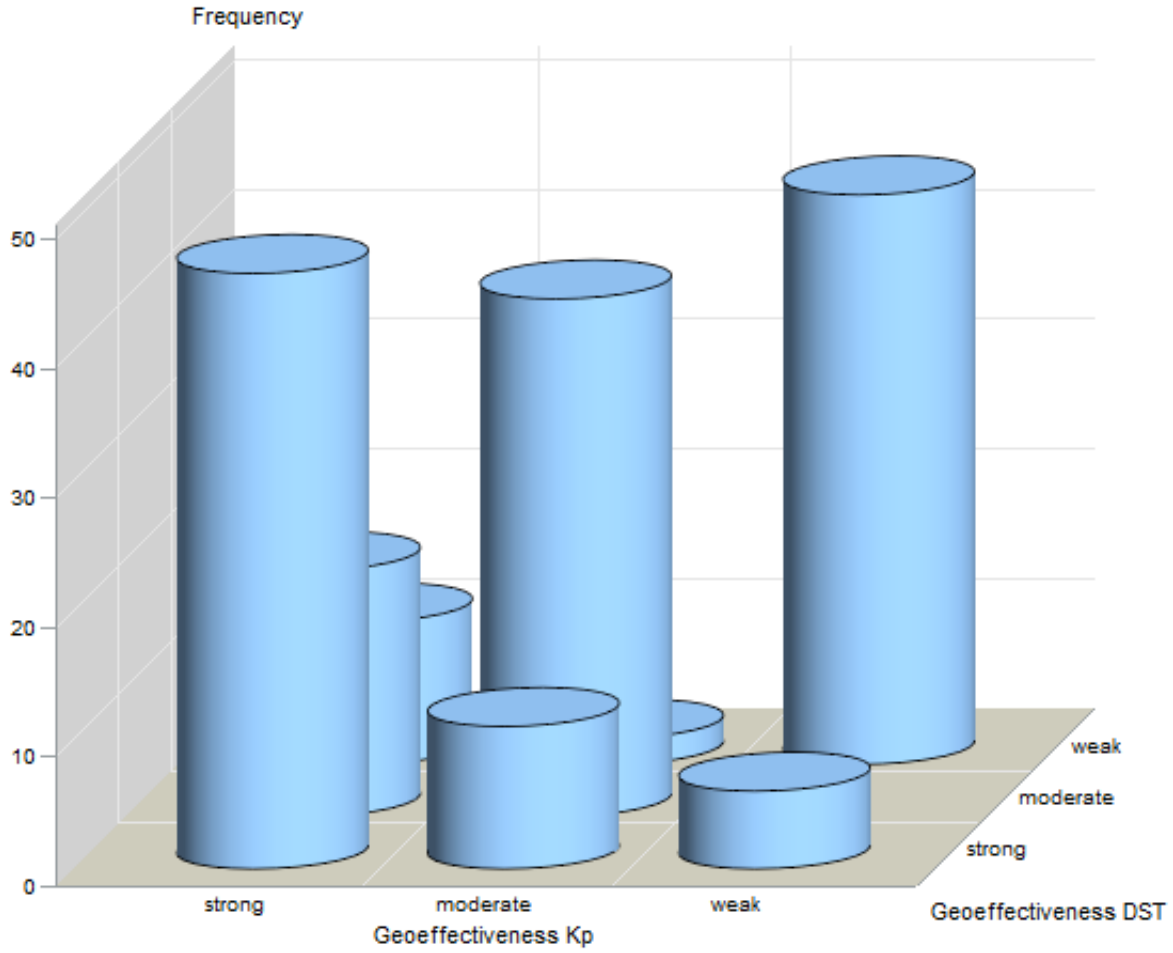


Figure 2.4: 3D graph of the contingency table comparing the classifications from both the DST and Kp.

Class Distribution of Geoeffectiveness			
	Strong	Moderate	Weak
Dst	59	63	57
Kp	53	76	50

Table 2.9: Number of ICME events that fall within in each category according to the different geomagnetic indices.

serve as the response variable for this work.

As noted by Kim et al. [54], the southward magnetic field component B_z[5] and the interplanetary electric field E_y (the interaction of B_z[6] and V_{sw} denoted by $E_y = [-V_{sw} * B_z]^{10^{-3}}$) yield the strongest linear relationship with the minimum DST. This trend is also true for this dataset (see Figure 2.5) with correlation coefficients of $r_{B_z} = 0.852$ and $r_{E_y} = -0.850$. These are almost identical to the correlation coefficients as found by the authors. Naturally, these predictor variables play an important role in determining dangerous geomagnetic storms [34] [51]. The next closest in magnitude is V_{sw} with a correlation coefficient of $r_{V_{sw}} = -0.406$.

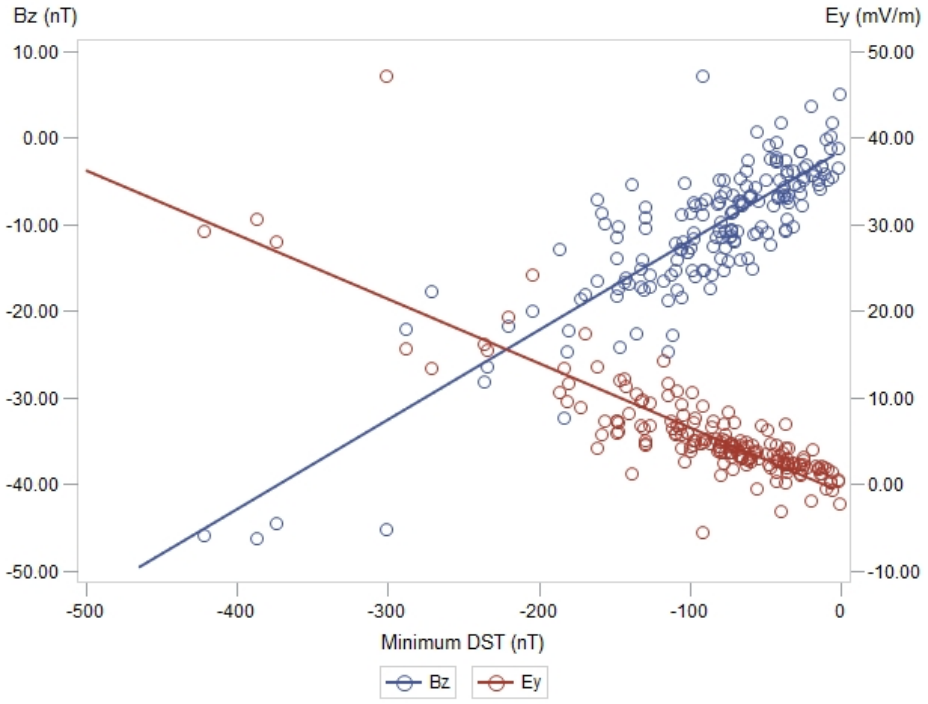


Figure 2.5: The B_z and E_y plotted against the minimum DST.

Another interesting predictor variable to explore is a CME's angular width (AW) given by the LASCO instrument aboard the SOHO satellite (see Figure 2.6).

Full halos are typically declared when $AW = 360\,^\circ$ and partial halos when $120\,^\circ \leq AW < 360\,^\circ$ [40]. Many studies have considered this characteristic in determining geoeffectiveness. In particular, Zhang, Dere, Howard, and Bothmer [94] as well as Srivastava and Venkatakrishnan [82] explored the impact of full halo CMEs. While both studies

[5]Measured by the Geocentric Solar Ecliptic Cartesian coordinates
[6]Measured by the Geocentric Solar Magnetospheric Cartesian coordinates

Figure 2.6: Box plot of the minimum DST segmented by CME halo classification. The category "1" represents an angular width of 360 ° and "0" otherwise.

indicate that full halo CMEs can cause major geomagnetic storms, the researchers concluded that not every full halo CME is dangerous. Their results are consistent with this dataset since full halo CMEs do not appear to severely alter the DST.

2.3.7 Implementation

Experiments for both the simulation study and predictive modeling on the CME dataset are conducted within the R environment for statistical computing [74] version 3.2.5. For the simulation study, application of RFs and RRFs are instituted using the *RRF* function in the **RRF** package [28] [29] [27], and application of CIRFs are instituted using the *cforest* function in the **party** package [46] [84] [83]. For assessing predictive performance on the CME dataset for the proposed CIGRRF and other popular classification techniques, the **caret** package [36] is used.

2.3.8 Simulation Study Set-up

Because of the biases associated with the Gini importance scheme, the construction of F in a GRRF will undoubtedly also be biased. However, this can be improved upon by using a CIGRRF for datasets where the predictor variables are a mixture of varying

scales of measurement and number of categories. To demonstrate this point, a simulation dataset is constructed. Similar to the power case simulation study by Strobl et al. [84], one predictor variable is made to be informative for the response while the others are randomly generated from different distributions. Compared to GRRFs, CIGRRFs should be able to detect the informative variable while also selecting the uninformative ones at a much lower frequency but at relatively the same rate, since it is less biased in its variable selection process. Ideally, with a truly unbiased model, the informative variable would be selected every time while the others are included in F at relatively the same frequency due to randomness. Since the CME dataset is a multiclass problem, the simulation study will also represent a multiclass problem. A summary of the simulated variables can be found in Table 2.10. $X_{01} - X_{03}$ are normally distributed for some mean with a standard deviation of one equal to $N(\mu, 1)$. $X_{04} - X_{06}$ are distributed as Poisson for some mean and standard deviation equal to $P(\lambda)$. $X_{08} - X_{10}$ as well as Y are distributed from a mulitnomial distribution with k classes equal to $M(k)$ where each class has an equal probability of occurring. The informative variable X_7 is a binary derived from Y with an indicator function. The sample size is set to $n = 250$ and the importance coefficient for both GRRF and CIGRRF is set to $\gamma = 1$, so that the penalization of each predictor variable is completely driven by the initial importance scores. This will provide a clear insight as to how biased GRRFs can be, and how much of this can be improved with the use of CIRFs. The number of trees for the initial calculation of the importance scores is set to 1,000 for both GRRF and CIGRRF for more stability [85]. All other arguments are left at their default settings. As done by Strobl et al. [84], 1,000 iterations are conducted. Each iteration generates a new dataset and new estimation of F for both the GRRF and CIGRRF.

2.3.9 Modeling on the CME Dataset

To investigate the predictive competitiveness of a CIRF using predictor variables selected by a CIGRRF, it is important to examine the performance of other classification models. This can be easily done with **caret**. This package allows for a streamlined user interface for applying a diverse set of predictive models from different packages with op-

Simulated Variables	Distribution	Type
Y	$\sim$ M(4)	Categorical
X_{01}	$\sim$ N(0,1)	Continuous
X_{02}	$\sim$ N(100,1)	Continuous
X_{03}	$\sim$ N(10000,1)	Continuous
X_{04}	$\sim$ P(0.5)	Discrete
X_{05}	$\sim$ P(1)	Discrete
X_{06}	$\sim$ P(2)	Discrete
X_{07}	$= \begin{cases} 1 & Y = \{1,2\} \\ 0 & Y = \{3,4\} \end{cases}$	Binary
X_{08}	$\sim$ M(4)	Categorical
X_{09}	$\sim$ M(10)	Categorical
X_{10}	$\sim$ M(20)	Categorical

Table 2.10: List of simulated variables for simulated study.

tions to perform various pre-processing, post-processing, resampling, and visualization techniques [56]. In addition, for those models that can perform variable importance estimation, the **caret** package can automatically extract these measures and rank them with a normalized importance score based on that model's strategy for establishing important predictor variables. Each model tested against the proposed approach can:

- Perform multiclass classification

- Calculate class probabilities

- Estimate the importance of each predictor variable with its own scheme

A summary of the 15 models also executed on the CME dataset are listed in Table 2.11.

Another advantage to using **caret** is the option to easily tune the parameters for a given algorithm or model by simply specifying a number for *tuneLength* in the *train* function. Each model has a predefined range of tuning values to search over proportional to *tuneLength*. The higher the *tuneLength*, the more searching executed. The number of tuning parameters ranges for each model. The final parameter settings are determined by those that deliver the best performance according to some metric. In this experimental set-up, *tuneLength* is left at the default value of three. To implement the proposed CIGRRF, a custom model is created within **caret**. The parameter tuned over for the CIGRRF is *mtry* (during the initial run of the CIRF and the following RRF). The

Model/Algorithm Name	**caret** Method
Bagged AdaBoost (ADA)	"AdaBag"
C5.0 Decision Tree and Rule-Based Model (C5.0)	"C5.0"
Classification and Regression Tree (CART)	"rpart"
Conditional Inference Random Forest (CIRF)	(see caption)
eXtreme Gradient Boosting (XGB)	"xgbLinear"
Flexible Discriminant Analysis (FDA)	"fda"
Lasso and Elastic Net (GLMNET)	"glmnet"
Nearest Shrunken Centroid (NSC)	"pam"
Neural Network (NN)	"nnet"
Partial Decision Trees Rule Learner (PART)	"PART"
Partial Least Squares (PLS)	"pls"
Penalized Multinomial Regression (PMR)	"multinom"
Random Forest (RF)	"rf"
RIPPER Rule Learner (RIPPER)	"JRip"
Stochastic Gradient Boosting (SGB)	"gbm"

Table 2.11: Other classification techniques implemented from the **caret** package. For the CIRF, **caret** calls the *cforest* function. Because this function can have different hyper-parameter settings, a custom model is created that specifies the unbiased option explicitly to ensure the unbiased implementation is used.

importance coefficient for the CIGRRF is set to $\gamma = 0.5$, since this provides a good balance between the number of predictor variables selected and the error rate in GRRFs [29]. As done in the simulation study, the default number of trees is increased from 500 to 1,000 in the preliminary run of the CIRF. As noted earlier, the final predictions are generated from a CIRF built on the predictor variables selected from the CIGRRF. The final execution of the CIRF is left at the default settings except for $mtry$, which is set to the number of predictor variables in F. A high-level sketch of the fit function for the custom model can be found via Algorithm 1.

2.3.10 Estimating Predictive Performance

As with any classification problem, it is imperative to estimate the error rate on an independently held-out sample of the data. However, this becomes difficult in situations where data is scarce. Hence, resampling approaches have been developed as a solution. Using k-fold cross-validation is a common strategy. Define the dataset $S = \{(x_i, y_i), i = 1, ..., n\}$. Specifically, split the dataset S into k near equal and disjoint sets such that $S_1, S_2, ..., S_k$. Let $S^{-k} = S - S_k$ and S_k be the training and test sets, respectively. Execute

Algorithm 1 Pseudo R Code–CIGRRF fit function in **caret** custom model.

```
procedure CIGRRF
    training ← cbind.data.frame(X, Y)
    for each parameter combination mtry = {2, 11, 20} and gamma = 0.5 do
        init.ctl ← cforest_unbiased(mtry = mtry, ntree = 1000) ▷ Setting parameters
        init.CIRF ← cforest(Y ∼ ., data = training, controls = init.ctl)
        imp ← varimp(init.CIRF, conditional = FALSE)
        impCIRF ← (imp - min(imp)) / (max(imp) - min(imp))          ▷ Normalizing
        lambda ← (1 - gamma) + (gamma * impCIRF)
        F ← RRF(Y ∼ ., data = training, mtry = mtry, coefReg = lambda)$feaSet
        reduced.training ← training[, F]        ▷ Only using chosen predictor variables
        final.ctl ← cforest_unbiased(mtry = length(F))      ▷ Setting parameters again
        final.CIRF ← cforest(Y ∼ ., data = reduced.training, controls = final.ctl)
    return final.CIRF
    end for
end procedure
```

a model on the first S^{-k} parts and produce predictions for the held-out part S_k. Repeat this procedure until each subset of S has been used as a test set exactly once. Averaging the error from each of these held-out parts gives a final estimate of the overall error rate for a model. This process yields roughly unbiased estimates of the true error rate but can be susceptible to high variance issues. To lower the variation of this estimator, it is common to perform the k-fold cross-validation procedure several times and then take the average. It has been shown that using repeated k-fold cross-validation performs well in estimating the true error rate for binary classification problems compared to other approaches such as .632+ bootstrapping [53]. However, when parameter tuning is involved, reporting the cross-validated estimate of error at a model's best parameter settings can be too optimistic; therefore, it is necessary to use nested cross-validation [90]. This procedure has two parts: an outer loop and inner loop of k-fold cross-validation. The outer folds represent the initial split of the data $S_1, S_2, ..., S_k$. The inner folds represents an internal k-fold cross-validation execution with training and tests folds constructed from S^{-k}. The purpose of the inner loop is the find the best parameters while the outer loop is to estimate predictive performance. Note that it is entirely possible to have different parameters chosen when evaluating the outer loop folds depending on the data split [70]. Given the small number of observations for the constructed CME dataset, 10 repeats of

10-fold (10×10) nested cross-validation are conducted to best determine the performance of each model.

For many of the previous studies in predicting geoeffective CMEs, the main performance metric utilized is accuracy. However, using only accuracy can have some pitfalls, especially when the class distributions are imbalanced. For instance, it is entirely possible for a model or algorithm to predict the majority class for all events in order to yield the highest accuracy, rendering the prediction useless for practical means. In addition, accuracy only tests the error rate based on a single threshold of the class probabilities (0.5) while other tools, such as Receiver Operating Characteristic curves [80], examine the proportion of actual positives classified correctly (true positive rate or recall) against the proportion of actual negatives that are incorrectly classified (false positive rate) across a variety of threshold points. In addition, it is independent from the prevalence or misclassification cost of a class [77]. Therefore, along with accuracy, this work calculates other metrics such as the area under the receiver operating characteristic curve (AUC). This metric has been shown to be a better one-number summary when evaluating different learning algorithms compared to accuracy [15] [49]. Hence, parameter combinations will be based on optimizing the AUC metric during each iteration in the 10×10 nested cross-validation. In addition to AUC and accuracy, other metrics including the area under the precision-recall curve (PRAUC) [75], LogLoss [38], and kappa statistic (kappa) [78] are also reported [77]. The PRAUC is similar to AUC, except it plots the proportion of predicted positives that are truly positive (precision) against the recall. LogLoss is a calibration score that seeks to calculate how close a probability prediction is to the true class designation [77]. Kappa [92] measures the accuracy of a model when juxtaposed against predictions made at random. Because the present work has a multiclass set-up, the "one-versus-all" strategy will be used for each of these traditionally binary performance metrics [56]. That is, each metric will be computed three times by creating three binary problems after a model makes its predictions:

- strong versus (moderate and weak)
- moderate versus (strong and weak)

- weak versus (strong and moderate)

These are then averaged to give a final estimate. For AUC, PRAUC, and kappa, stronger classifiers will have values closer to one. For LogLoss, the smaller the value the better. Additionally, the number of non-zero predictor variables determined by each model's variable importance scheme are averaged from each iteration within the nested cross-validation process. In this way, the amount of parsimony for each model can be quantified.

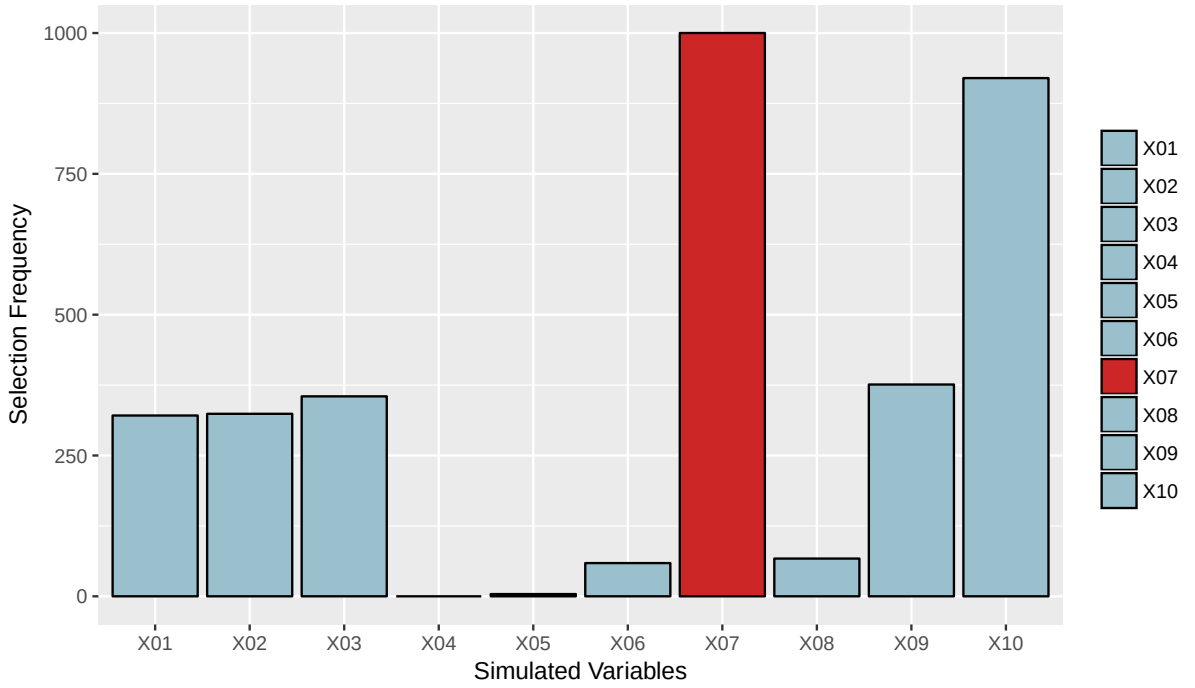
Finally, it is important to assess whether the proposed CIGRRF performs significantly better than the other classification methods tested. To accomplish this, this work implements the corrected repeated k-fold cross-validation test [14], also known as the corrected t-test for repeated cross-validation [77], on the population of performance metrics (100 estimates from the 10×10 nested cross-validation). This test has been shown to produce acceptable Type I error, low Type II error, and good replicability for comparing two models on classification tasks [14] [77].

2.4 Results

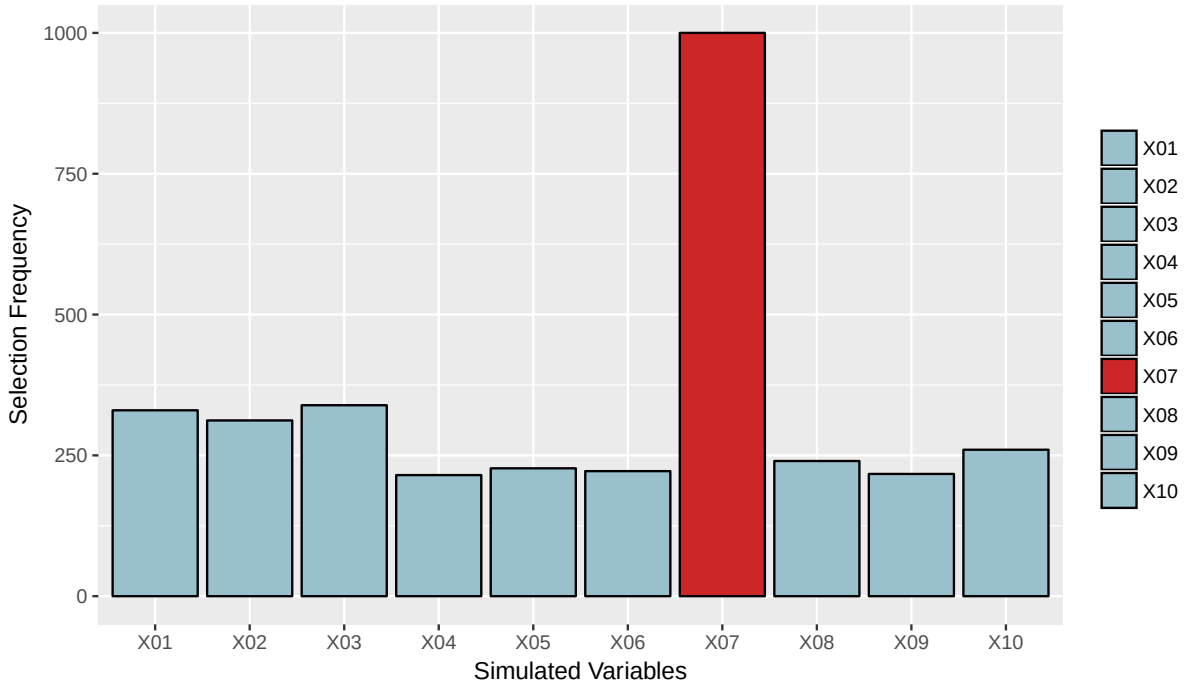
2.4.1 Simulation Study Results

Figure 2.7 displays the variable selection frequencies for both GRRF and CIGRRF from the 1,000 iterations. Both variable selection schemes are able to select the informative variable X_{07} at every iteration and produce approximately the same number of predictor variables at each iteration on average.

However, as expected, the GRRF includes the uninformative continuous variables $(X_{01} - X_{03})$ and those with many categories (X_{09}, X_{10}) at a much higher frequency in F compared to the others. On the other hand, the CIGRRF shows much less bias and selects the uninformative ones at relatively the same rate, no matter the type. It is evident that in comparison to the GRRF, the CIGRRF is a much more reliable variable selection process for data with different scales of measurement. This results in cleaner sets of relevant and non-redundant predictor variables in F for interpretation and predictive



(a) $\overline{F} = 3.43$



(b) $\overline{F} = 3.36$

Figure 2.7: Simulated variable selection frequency for the (a) GRRF and the (b) CIGRRF. The red bar indicates the informative variable X_{07}. The average size of F after each execution over the course of the 1,000 iterations is given by $\overline{F}$.

purposes.

2.4.2 Predictive Performance on the CME Dataset

Table 2.12 displays the mean values of each metric and $\overline{F}$ according to the 10×10 nested cross-validation procedure. Bold and italics values indicate the model with the best performance. Rankings based on the performance of the CIRF with CIGRRF variable selection (CIGRRF-CIRF) is given. The dagger symbol "†" indicates when statistical differences between the CIGRRF and the other classification models are found at the conventional 0.05 significance level.

The proposed approach delivers the second best value in the probability based metrics AUC and LogLoss while performing in the top five for PRAUC, accuracy, and kappa. Moreover, it performs significantly better than eight of the 15 competing classification models for the important AUC metric. Naturally, it functions on par with other bagging and boosting tree ensembles (ADA, CIRF, RF, SGB, and XGB). However, the CIGRRF-CIRF needs the fewest predictor variables on average on this dataset to make future predictions. In addition, it is not outperformed with statistical difference according to any of the metrics. Hence, it exhibits a good balance between parsimony and predictive power by selecting the fewest number of predictor variables while also delivering competitive performance.

Interestingly enough, the simple CART model gives the best accuracy and kappa measures with relatively few predictor variables. However, the CIGRRF-CIRF performs significantly better in terms of AUC and PRAUC. This suggests that CART does well at one particular classification threshold, but not so much when this is varied, especially when trying to strike a balance between precision and recall (low PRAUC value relatively to all the others). Additional reasoning for using the CIGRRF-CIRF as opposed to CART is in regards to the importance scores. For CART, these are based on the aforementioned Gini index, which is biased as shown in this study via simulations.

When comparing to the CIRF (which yields the probability based metrics), the posited modification performs statistically similarly while considering far fewer predictor variables on average. This indicates that the CIGRRF can eradicate the less important predic-

tor variables to provide a much sparser solution. Overall, based on these results, the CIGRRF-CIRF should be the preferred classifier for this dataset.

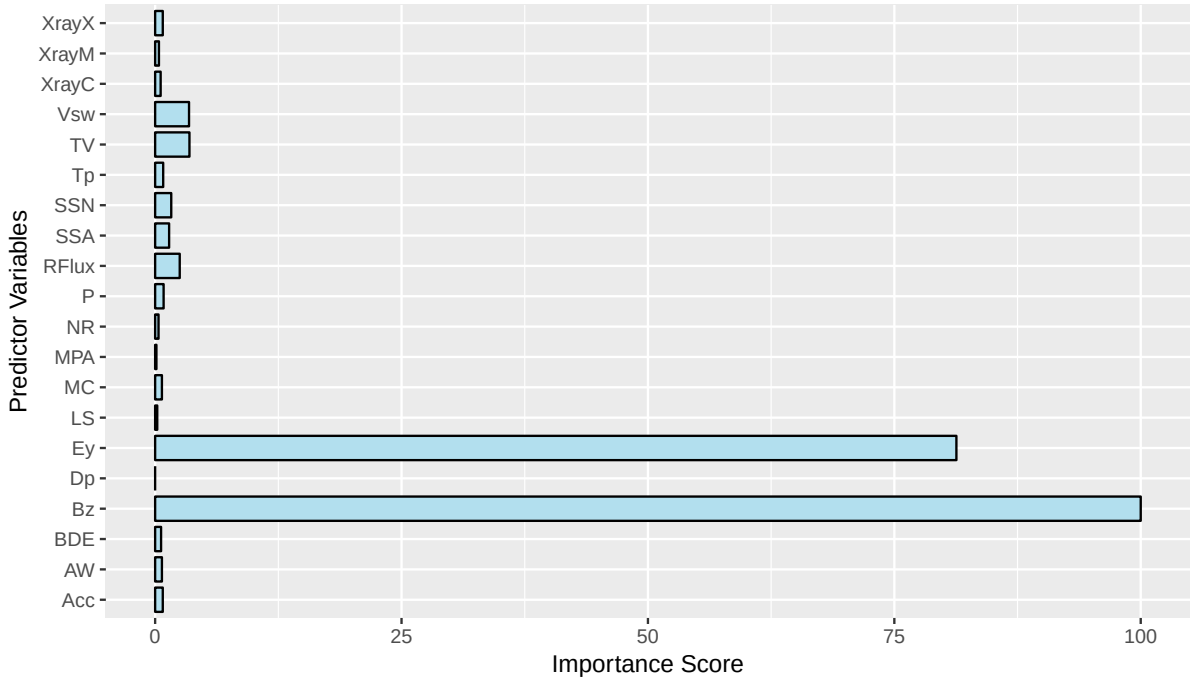
Model	AUC	PRAUC	LogLoss	Accuracy	Kappa	$\overline{F}$
ADA	0.8649	0.6101	0.5928	0.7393	0.6079	15.04
C5.0	0.8358†	0.5710	0.6493†	0.6749	0.5121	14.34
CART	0.8311†	0.1257†	0.4340	***0.7535***	***0.6294***	6.33
XGB	0.8485	0.6236	0.6049†	0.6928	0.5385	19.51
FDA	0.8422	0.5959	0.6044†	0.7056	0.5571	3.55
GLMNET	0.8485	0.6222	0.4703	0.7060	0.5585	8.10
NSC	0.6700†	0.4430†	0.6532†	0.4937†	0.2393†	15.56
NN	0.5230†	0.2066†	0.6611†	0.3564†	0.0242†	20.00
PART	0.7358†	0.2940†	2.4330†	0.6306†	0.4453†	15.57
PLS	0.6218†	0.4059†	0.6255†	0.4336†	0.1443†	20.00
PMR	0.8090†	0.5709	0.6389†	0.6566†	0.4844†	20.00
RF	0.8696	0.6437	0.4151	0.7365	0.6036	20.00
RIPPER	0.8098†	0.1834†	0.5968	0.7123	0.5676	2.72
SGB	0.8599	0.6342	0.4860	0.7117	0.5668	14.78
CIRF	***0.8800***	***0.6589***	***0.4054***	0.7325	0.5984	19.79
CIGRRF-CIRF	0.8726	0.6254	0.4128	0.7185	0.5776	2.50
Rank	(2)	(4)	(2)	(5)	(5)	(1)

Table 2.12: Predictive performance of the proposed approach as well as the comparison models.

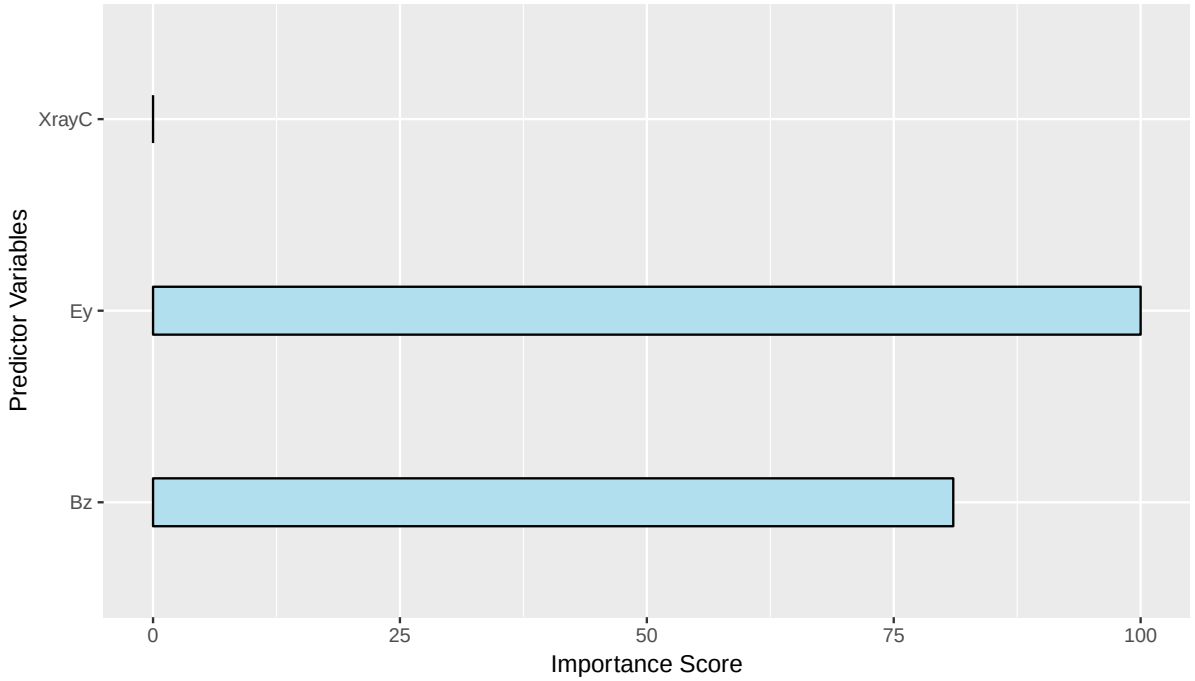
Figure 2.8 displays the min-max normalized variable importance scores based on the conditional[7] permutation scheme for both the normal run of the CIRF and that which uses the predictor variables selected from the proposed CIGRRF. The blue bar indicates the importance of each predictor variable with values ranging from zero to 100 such that 100 signifies the most important. These are calculated at their best *mtry* settings as determined by an ordinary run of 10-fold cross-validation on all the data. This final training of each model is necessary so that the variable importance scores are constructed utilizing all of the data.

The two most important predictor variables in the full CIRF (B_z and E_y) are also included in the parsimonious version. Given the strong relationship between these predictor variables and the DST, their contributions towards prediction makes sense. However, note that they are in different orders. The CIGRRF dictates that E_y is most important

[7]Using the conditional variant here is feasible since this only needs to be executed once after each is trained on all of the data.



(a) CIRF



(b) CIGRRF-CIRF

Figure 2.8: Normalized conditional permutation variable importance scores for the two CIRFs when (a) using all the predictor variables (b) using only those selected by the CIGRRF.

while the CIRF indicates B_z. This could be due to using the default value of *ntree* at the final execution of the CIRF. It is likely that increasing the number of trees will yield more consistency between these two models [85]. Regardless, the higher value placed on these ICME predictor variables and lower values on those such as D_p and T_p in determining geomagnetic storm intensity is consistent with other literature (see [93] [34] [51] [54] and references therein). This helps confirm the integrity of the dataset constructed as well as the ability for the CIGRRF to choose the important predictor variables E_y and B_z and exclude the less important ones like D_p and T_p. In addition, AW is also not included in F for the CIGRRF and is scored very low in CIRF. Similar to Figure 2.6, AW does not seem to greatly impact prediction in the presence of the other information.

2.5 Discussion

In this work, a modified version of RRFs is explored for classifying geoeffective CMEs. Based upon the algorithm for GRRFs, which incorporates a preliminary RF's variable importance scores to help guide the variable selection process in a RRF, the proposed CIGRRF replaces the initial RF with a CIRF. Doing so allows for the unconditional permutation variable importance scores from an unbiased tree ensemble to penalize the information gain for each predictor variable, as opposed to using the Gini importance scheme which has been shown to be biased. This modification helps alleviate some of the glaring favoritism towards continuous predictor variables and ones with many categories in GRRFs as shown via a simulation study. In addition, this change can lead to smaller predictor variable subsets produced by RRFs for a more parsimonious prediction. These subsets are introduced into a CIRF since these have many advantages for interpretation (e.g. unbiasedness, conditional variable importance). The CIGRRF is able to identify the most prominent drivers for severe geomagnetic storms as regarded by the literature while being able to eliminate much of the noise via variable selection. Not only do CIGRRFs have interpretative benefits, combining these with a CIRF as the classifier can perform competitively against other sophisticated ensemble models as seen on this dataset.

As mentioned before, the unconditional permutation variable importance scheme can

exhibit some issues identifying influential predictor variables in highly correlated situations. One could potentially use the conditional variable importance framework to create the coefficient of regularization for a more reliable initial variable importance estimation. However, this process is computationally demanding, especially for larger datasets, which may make this approach potentially infeasible in practice. Perhaps an approximation could lead to faster calculations, thereby, leading to an improvement for CIGRRFs. In addition, the predictor variable selected by CIGRRFs do not necessarily need to be fed into a CIRF. Other models can be used at this step. However, since these tree models have some favorable properties, they seem to be logical candidates. Moreover, analyzing these variable importance scores via the permutation approach offers an alternative and model based way to evaluate useful predictor variables that can provide additional information for those concerned with space weather.

Since one dataset and one importance coefficient value are considered for prediction, a more comprehensive assessment on a variety of datasets against possibly more classification techniques is needed to make a full evaluation of CIGRRFs and their predictive capabilities compared to other methods. As found by Deng and Runger [29], better predictive performance is achieved for lower values of γ in general. Since lower values of γ lead to larger sets of F, researchers and practitioners may want to tune this in order to reach a desired balance between predictive performance and the number of predictor variables selected, depending on his or her goal. Given the intimate connection between GRRFs and CIGRRFs, it is reasonable that the conclusions are the same for tuning γ in CIGRRFs. In addition, only one simulation study is performed in this work. A more thorough treatment of CIGRRFs via more extensive simulation studies are needed to fully understand the behaviors of the proposed approach.

In this work, care is taken to ensure a realistic dataset is constructed from a variety of data sources to provide a natural prediction scenario for these models in term of classifying geoeffective CMEs. Future work could consist of exploring a larger number of CMEs and re-assessing the benefits of CIGRRFs for geomagnetic storm prediction. In addition, no cost matrix is considered here. Being able to incorporate the potential costs

of making incorrect predictions can help calibrate these models appropriately. This can easily be implemented into CIRFs and other tree based methods. Furthermore, because the dataset is constructed from data obtained near the Sun and from the important interplanetary parameters collected closer to Earth, predictions can only be made a few hours out prior to the geomagnetic storm. Building frameworks to make more timely predictions is of valuable interest [54].

2.6 Conclusion

Due to the potentially catastrophic consequences of CMEs on the global business environment, it is a necessity to construct highly accurate yet interpretive algorithmic processes. Since it is not feasible for business entities to take extreme precautionary measures for every CME event that approaches Earth, these processes must help dictate when those extreme measures are warranted. By using the information from a CIRF in a RRF, more reliable penalties can be applied to each predictor variable. This is especially important for datasets with a wide range of measurements such as for CME datasets. Not only does this lead to more unbiased variable selection, but it can provide a simpler model without sacrificing much predictive power. Moreover, the ability to create a more parsimonious approach to CIRFs can further help practitioners and researchers investigate important space weather information using the innovate approaches of tree ensembles. Continuing to investigate these phenomena cannot be overemphasized as society is becoming increasingly dependent on technology. Thankfully, modern capabilities allow for the collection of vast amounts of data regarding geomagnetic storms. Hence, utilizing more data-driven approaches from both machine learning and statistical fields on these data will continue to aid in the important task of space weather prediction.

2.7 References

[1] National Aeronautics and Space Administration. CME week: The difference between flares and CMEs. Retrieved from https://www.nasa.gov/content/goddard/the-difference-between-flares-and-cmes [accessed: 2016-11-18].

[2] National Aeronautics and Space Administration. Coronal mass ejections. Retrieved from http://helios.gsfc.nasa.gov/cme.html [accessed: 2016-05-16].

[3] National Aeronautics and Space Administration's Goddard Space Flight Center. Rattling earth's force field. Retrieved from https://svs.gsfc.nasa.gov/10954 [accessed: 2016-07-27].

[4] Joe Allen, Herb Sauer, Lou Frank, and Patricia Reiff. Effects of the March 1989 solar activity. *Eos, Transactions American Geophysical Union*, 70(46):1479–1488, 1989.

[5] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

[6] Lloyd's , Atmospheric and Environmental Research Inc. Solar storm risk to the north american electrical grid, 2013. Retrieved from https://www.lloyds.com/~/media/lloyds/reports/emerging%20risk%20reports/solar%20storm%20risk%20to%20the%20north%20american%20electric%20grid.pdf [accessed: 2016-09-06].

[7] DN Baker, X Li, A Pulkkinen, CM Ngwira, ML Mays, AB Galvin, and KDC Simunac. A major solar eruptive event in July 2012: Defining extreme space weather scenarios. *Space Weather*, 11(10):585–591, 2013.

[8] J Bartels, NH Heck, and HF Johnston. The three-hour-range index measuring geomagnetic activity. *Terrestrial Magnetism and Atmospheric Electricity*, 44(4):411–454, 1939.

[9] Aon Benfield. Geomagnetic storms, 2013. Retrieved from http://thoughtleadership.aonbenfield.com/Documents/201301_geomagnetic_storms.pdf [accessed: 2016-09-06].

[10] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[11] DH Boteler. The super storms of August/September 1859 and their effects on the telegraph system. *Advances in Space Research*, 38(2):159–172, 2006.

[12] DH Boteler, RJ Pirjola, and H Nevanlinna. The effects of geomagnetic disturbances on electrical systems at the Earth's surface. *Advances in Space Research*, 22(1):17–27, 1998.

[13] Volker Bothmer and Rainer Schwenn. The interplanetary and solar causes of major geomagnetic storms. *Journal of Geomagnetism and Geoelectricity*, 47(11):1127–1132, 1995.

[14] Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in Knowledge Discovery and Data Mining*, pages 3–12. Springer, 2004.

[15] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[16] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[17] Leo Breiman. Out-of-bag estimation. Technical report, Citeseer, 1996.

[18] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[19] Leo Breiman and Adele Cutler. Random forests. Retrieved from `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro` [accessed: 2016-05-25].

[20] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

[21] GE Brueckner, J-P Delaboudiniere, RA Howard, SE Paswaters, OC St Cyr, R Schwenn, P Lamy, GM Simnett, B Thompson, and D Wang. Geomagnetic storms caused by coronal mass ejections (CMEs): March 1996 through June 1997. *Geophysical Research Letters*, 25(15):3019–3022, 1998.

[22] LF Burlaga, L Klein, NR Sheeley, DJ Michels, RA Howard, MJ Koomen, R Schwenn, and H Rosenbauer. A magnetic cloud and a coronal mass ejection. *Geophysical Research Letters*, 9(12):1317–1320, 1982.

[23] HV Cane and IG Richardson. Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, 108(A4), 2003.

[24] Richard C Carrington. Description of a singular appearance seen in the Sun on September 1, 1859. *Monthly Notices of the Royal Astronomical Society*, 20:13–15, 1859.

[25] Joseph M Caswell and Nicolas Rouleau. Simple binary prediction of daily storm-level geomagnetic activity with solar winds and potential relevance for cerebral function. *International Letters of Chemistry, Physics and Astronomy*, 17, 2014.

[26] EW Cliver. The 1859 space weather event: Then and now. *Advances in Space Research*, 38(2):119–129, 2006.

[27] Houtao Deng. Guided random forest in the RRF package. *arXiv preprint arXiv:1306.0237*, 2013.

[28] Houtao Deng and George Runger. Feature selection via regularized trees. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

[29] Houtao Deng and George Runger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483–3489, 2013.

[30] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3, 2006.

[31] S Dimitrova, I Stoilova, K Georgieva, T Taseva, M Jordanova, and D Maslarov. Solar and geomagnetic activity and acute myocardial infarction morbidity and mortality. *Vascular Diseases*, 8:9, 2009.

[32] M Dryer, Z Smith, CD Fry, W Sun, CS Deehr, and S-I Akasofu. Real-time shock arrival predictions during the Halloween 2003 epoch. *Space Weather*, 2(9), 2004.

[33] James W Dungey. Interplanetary magnetic field and the auroral zones. *Physical Review Letters*, 6(2):47, 1961.

[34] E Echer, WD Gonzalez, and BT Tsurutani. Interplanetary conditions leading to superintense geomagnetic storms (DST <= -250 nT) during solar cycle 23. *Geophysical Research Letters*, 35(6), 2008.

[35] Donald H Fairfield and LJ Cahill. Transition region magnetic field and polar magnetic disturbances. *Journal of Geophysical Research*, 71(1):155–169, 1966.

[36] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. *caret: Classification and Regression Training*, 2016. R package version 6.0-68.

[37] Walter D Gonzalez and Bruce T Tsurutani. Criteria of interplanetary parameters causing intense magnetic storms (DST <-100 nT). *Planetary and Space Science*, 35(9):1101–1109, 1987.

[38] Irving John Good. Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *The British Journal for the Philosophy of Science*, 19(2):123–143, 1968.

[39] N Gopalswamy, S Yashiro, G Michalek, G Stenborg, A Vourlidas, S Freeland, and R Howard. The SOHO/LASCO CME catalog. *Earth, Moon, and Planets*, 104(1-4):295–313, 2009.

[40] Nat Gopalswamy. Halo coronal mass ejections and geomagnetic storms. *Earth, Planets and Space*, 61(5):595–597, 2009.

[41] Nat Gopalswamy, Alejandro Lara, Seiji Yashiro, Mike L Kaiser, and Russell A Howard. Predicting the 1-AU arrival times of coronal mass ejections. *Journal of Geophysical Research: Space Physics (1978–2012)*, 106(A12):29207–29217, 2001.

[42] John T Gosling, DN Baker, SJ Bame, WC Feldman, RD Zwickl, and EJ Smith. Bidirectional solar wind electron heat flux events. *Journal of Geophysical Research: Space Physics (1978–2012)*, 92(A8):8519–8535, 1987.

[43] JT Gosling, SJ Bame, DJ McComas, and JL Phillips. Coronal mass ejections and large geomagnetic storms. *Geophysical Research Letters*, 17(7):901–904, 1990.

[44] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[45] Douglas M Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

[46] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[47] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

[48] Tim Howard. *Coronal Mass Ejections: An Introduction*, volume 376. Springer Science & Business Media, 2011.

[49] Jin Huang and Charles X Ling. Using AUC and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 17(3):299–310, 2005.

[50] KEJ Huttunen, R Schwenn, V Bothmer, and HEJ Koskinen. Properties and geoeffectiveness of magnetic clouds in the rising, maximum and early declining phases of solar cycle 23. In *Annales Geophysicae*, volume 23, pages 1–17, 2005.

[51] Eun-Young Ji, Y-J Moon, K-H Kim, and D-H Lee. Statistical comparison of interplanetary conditions causing intense geomagnetic storms (DST <= -100 nT). *Journal of Geophysical Research: Space Physics (1978–2012)*, 115(A10), 2010.

[52] JG Kappernman and Vernon D Albertson. Bracing for the geomagnetic storms. *Spectrum, IEEE*, 27(3):27–33, 1990.

[53] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

[54] R-S Kim, Y-J Moon, N Gopalswamy, Y-D Park, and Y-H Kim. Two-step forecast of geomagnetic storm using coronal mass ejection and solar wind condition. *Space Weather*, 12(4):246–256, 2014.

[55] JH King and NE Papitashvili. Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. *Journal of Geophysical Research: Space Physics (1978–2012)*, 110(A2), 2005.

[56] Max Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[57] Alejandro Lara, Nat Gopalswamy, Hong Xie, Eduardo Mendoza-Torres, Román Pérez-Eríquez, and Gregory Michalek. Are halo coronal mass ejections special events? *Journal of Geophysical Research: Space Physics*, 111(A6), 2006.

[58] Roger J Lewis. An introduction to classification and regression tree (CART) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, pages 1–14, 2000.

[59] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[60] CA Loewe and GW Prölss. Classification and mean behavior of magnetic storms. *Journal of Geophysical Research: Space Physics (1978–2012)*, 102(A7):14209–14213, 1997.

[61] RM MacQueen. Coronal transients: A summary. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 297(1433):605–620, 1980.

[62] Denise McManus, Houston Carr, and Benjamin Adams. Wireless on the precipice: The 14th century revisited. *Communications of the ACM*, 54(6):138–143, 2011.

[63] National Oceanic and Atmospheric Administration. Aurora. Retrieved from `http://www.swpc.noaa.gov/phenomena/aurora` [accessed: 2016-05-16].

[64] National Oceanic and Atmospheric Administration. Coronal mass ejections. Retrieved from `http://www.swpc.noaa.gov/phenomena/coronal-mass-ejections` [accessed: 2016-05-16].

[65] National Oceanic and Atmospheric Administration. Earth's magnetosphere. Retrieved from `http://www.swpc.noaa.gov/phenomena/earths-magnetosphere` [accessed: 2016-05-16].

[66] National Oceanic and Atmospheric Administration. Index of /pub/warehouse. Retrieved from `ftp://ftp.swpc.noaa.gov/pub/warehouse` [accessed: 2015-04-15].

[67] National Oceanic and Atmospheric Administration. NOAA space weather scale. Retrieved from `http://www.swpc.noaa.gov/noaa-scales-explanation` [accessed: 2015-04-15].

[68] Sten Odenwald, James Green, and William Taylor. Forecasting the impact of an 1859-calibre superstorm on satellite resources. *Advances in Space Research*, 38(2):280–297, 2006.

[69] Sten F Odenwald and James L Green. Bracing for a solor superstorm. *Scientific American*, 299(2):80–87, 2008.

[70] Christian Petersohn. *Temporal Video Segmentation*. Jörg Vogt Verlag, 2010.

[71] George Bartlett Prescott. *History, Theory, and Practice of the Electric Telegraph*. Frank Jones, 1866.

[72] Antti Pulkkinen, Sture Lindahl, Ari Viljanen, and Risto Pirjola. Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system. *Space Weather*, 3(8), 2005.

[73] R Qahwaji, Tufan Colak, M Al-Omari, and S Ipson. Automated prediction of CMEs using machine learning of CME–flare associations. *Solar Physics*, 248(2):471–483, 2008.

[74] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[75] Vijay Raghavan, Peter Bollmann, and Gwang S Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

[76] IG Richardson and HV Cane. Near-Earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009): Catalog and summary of properties. *Solar Physics*, 264(1):189–237, 2010.

[77] Guzman Santafe, Iñaki Inza, and Jose A Lozano. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508, 2015.

[78] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1015–1021. Springer, 2006.

[79] Space Studies Board and others. *Severe Space Weather Events: Understanding Societal and Economic Impacts: A Workshop Report*. National Academies Press, 2008.

[80] Kent A Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 160–163. Morgan Kaufmann Publishers Inc., 1989.

[81] N Srivastava. A logistic regression model for predicting the occurrence of intense geomagnetic storms. In *Annales Geophysicae*, volume 23, pages 2969–2974, 2005.

[82] Nandita Srivastava and P Venkatakrishnan. Solar and interplanetary sources of major geomagnetic storms during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, 109(A10), 2004.

[83] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):1, 2008.

[84] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

[85] Carolin Strobl, Torsten Hothorn, and Achim Zeileis. Party on! A new, conditional variable importance measure for random forests available in the party package, 2009. Retrieved from `https://journal.r-project.org/archive/2009-2/RJournal_2009-2_Strobl~et~al.pdf` [accessed: 2016-09-06].

[86] Masahisa Sugiura. Hourly values of equatorial DST for the IGY. *Ann. Int. Geophys. Yr.*, 35, 1964.

[87] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[88] Bruce T Tsurutani and Walter D Gonzalez. The interplanetary causes of magnetic storms: A review. *Washington DC American Geophysical Union Geophysical Monograph Series*, 98:77–89, 1997.

[89] Jean Uwamahoro, Lee-Anne McKinnell, and John Bosco Habarulema. Estimating the geoeffectiveness of halo CMEs from associated solar and IP parameters using neural networks. *Annales Geophysicae-Atmospheres Hydrospheres and Space Sciences*, 30(6):963, 2012.

[90] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):1, 2006.

[91] YM Wang, PZ Ye, S Wang, GP Zhou, and JX Wang. A statistical study on the geoeffectiveness of Earth-directed coronal mass ejections from March 1997 to December 2000. *Journal of Geophysical Research: Space Physics (1978–2012)*, 107(A11):SSH–2, 2002.

[92] Ian H Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[93] Yu I Yermolaev, M Yu Yermolaev, IG Lodkina, and NS Nikolaeva. Statistical investigation of heliospheric conditions resulting in magnetic storms: 2. *Cosmic Research*, 45(6):461–470, 2007.

[94] J Zhang, KP Dere, RA Howard, and V Bothmer. Identification of solar sources of major geomagnetic storms between 1996 and 2000. *The Astrophysical Journal*, 582(1):520, 2003.

[95] Qifeng Zhou, Wencai Hong, Linkai Luo, and Fan Yang. Gene selection using random forest and proximity differences criterion on DNA microarray data. *Journal of Convergence Information Technology*, 5(6):161–170, 2010.

3

A TWO-STAGE META-LEARNING FRAMEWORK FOR PREDICTING GEOMAGNETIC STORMS

3.1 Introduction

3.1.1 Geomagnetic Storms

On September 16, 2016, the Business Insider published an article discussing the dangers of the space weather events known as solar storms [56]. These solar storms, or more technically called in the article as CMEs (also known as ICMEs as they propagate away from the Sun), are large masses of particles from the Sun released with incredible force. When these expulsions collide with Earth, they can strain and manipulate the protective magnetic field lines, leaving parts of Earth exposed to harmful solar material and sparking geomagnetic storms. If strong enough, these storms can overload power grid infrastructures by unleashing their overwhelming energy into electric systems on Earth. Not only is this a problem in the electricity sector, but also these storms can cause disruptions in telecommunication operations and damage satellites. Details regarding this in more depth can be found in a variety of sources [73] [36] [52] [47]. Within the Business Insider article, besides outlining the general peril of such phenomena, the authors acknowledge that the risk for Earth is not uniform. In other words, some areas are more at risk than others, such as Minnesota. This is due to its complicated geology and closer geographical location to the poles, which are more susceptible to the adverse effects from geomagnetic storms. Hence, companies such as Minnesota Power rely on alerts from the NOAA Space Weather Prediction Center to prepare for an impending solar storm. The article concludes on a somber note:

"While the electric industry catches up to the threat posed by the Sun, the best we can

do is hope our planet dodges the next solar storm – one barely missed us in July 2012 – and try to prepare for the worst." [56]

This is just one of countless articles purporting the dangers of geomagnetic storms and their potential, cataclysmic impact to electrical grids and telecommunications in recent years. While this may seem dismal, researchers have actively been involved in predicting the severity of these geomagnetic storms before they interact with Earth. NOAA has been at the forefront of predicting various space weather events, including CMEs, and posting their alerts and information online. Thanks to technological advancements and the launching of satellites such as the SOHO, more data are being collected about CMEs, which invites the opportunity for more data-driven analyses and predictions.

One of the major challenges with forecasting geomagnetic storms is the timeliness of the predictions. Because of interactions with the solar wind through the heliosphere, much of a CME's composition can be changed during its one to four day [75] passage away from the Sun [36]. Hence, using only the data recorded about a CME at its initial launch to make predictions as to its geoeffectiveness, or estimated impact on Earth, can be poor [40]. The most accurate information regarding its impact can be taken from satellites like the ACE, which records interplanetary information (IPI) such as the solar wind condition and its magnetic orientation at the L1 Lagrangian point (about 1.5 million kilometers from the Earth). While forecasts from this data source are quite good, it leaves only hours of lead time, or time available to make preparations on Earth before the associated shock arrives. To combat this issue, works by Valach, Bochníček, Hejda, and Revallo [86] and by Kim, Moon, Gopalswamy, Park, and Kim [40] offer a two-step approach to utilize both types of data to make a more informed decision. Another challenge is that only a small percentage of these events is directed towards Earth; thus, the majority of CMEs never cause any issues. Given the detrimental consequences, availability of data, and time constraints, sophisticated learning strategies (e.g. meta-learning) should be utilized to predict geomagnetic storms.

3.1.2 Meta-Learning

Brazdil, Carrier, Soares, and Vilalta [8] defined meta-learning (a phrase conceived by

Chan and Stolfo [15]) as "the process of invoking a learning algorithm to obtain knowledge concerning the behavior of a machine learning or data mining process" (terminology section). In addition, Vilalta and Drissi [89] explained that meta-learning "studies how learning systems can increase in efficiency through experience" and that "the goal is to understand how learning itself can become flexible according the domain or task under study" (p. 77). Essentially, the ultimate goal is to gain meta-knowledge, or information about the learning process. The obtained meta-knowledge can be leveraged for a variety of applications such as model selection, parameter optimization, and model combination [8]. This quality of self-adaptation becomes advantageous as practitioners must constantly recalibrate models to accommodate an ever-increasing flow of data about their customers, competitors, and business environment.

Being able to learn at the meta-level as opposed to the base-level has its dominance in terms of bias. Brazdil et al. [8] designate bias as "any preference for choosing one hypothesis explaining the data over other (equally acceptable) hypotheses" (p. 3). When implementing a learner (statistical model, machine learning algorithm, prediction technique, etc.), a fixed bias to choose a particular outcome is associated with that specific model. However, using a meta-learner allows for an adaptive selection of a model or a subset of models with the correct bias for a particular problem [89] [1]. In addition, learning only at the base-level inhibits the accumulation of meta-knowledge through experience about a business problem, therefore, leading to a loss of generality when training data from other sources of similar business tasks [8].

One type of meta-learning is the "meta-learner of base-learners" method [89] (p. 82). Known also as stacked generalization [92], this consists of using a set of learners at the base-level (base-learners) as inputs for another set of learner(s) at the meta-level (meta-learner(s)) [1]. It is considered a type of meta-learning because of the transfer of information about the base-level predictions to the meta-level [89]. The ability to learn from the base-learners helps capitalize on the strengths of each base-learner for higher accuracy at the meta-level. Hence, stacked generalization will likely perform at least as good, if not better, than the best base-learner for a given problem [9] [87]. This meta-

learning approach has been the backbone in highly complex problems such as financial fraud [1], bank failure [83], and user ratings in the famous Netflix Prize competition [72]. However, as discussed by Vilalta and Drissi [89], a weakness derived from using such an approach is that while each base-learner exhibits a fixed-form bias, so does the meta-learner. Hence, no dynamic bias selection is taking place at the meta-level. In addition, as with other ensemble approaches, stacked generalization is often viewed as a "black-art" [92] so making any sort of interpretations about which predictor variables are most important is typically lost. This work attempts to alleviate these issues via parameter tuning, using a sparse meta-learner, and offering a simple way to gain insight at the predictor variables in the dataset being studied.

3.1.3 The Purpose of This Work

CMEs can present a substantial threat for a variety of business entities. Because of the rarity of CMEs colliding with Earth, it is not practical for firms to shut down power grids and telecommunication operations every time a CME is detected. At the same time, completely ignoring these phenomena can lead to dire situations. Therefore, it is necessary to find accurate methods to predict the potential impact of an impending CME. Motivated by multi-step approaches, this work employs a two-stage framework to increase lead time and predictive power. In stage one, using a recently proposed random forest variant [47] to analyze the initial characteristics of a CME, preliminary probability estimates as to the geoeffectiveness of a CME based on the data available at its inception are made. Then, after incorporating the important IPI as the CME approaches Earth, stage two implements stacked generalization to effectively exploit the biases given by a set of base-learners. That is, the predictions from a variety of models as well as the probability estimates from stage one are collated as inputs for a meta-learner. For the respective meta-learner, a penalized quantile regression model [61] is implemented to effectively estimate the infrequent but dangerous geomagnetic storms as well as promote sparse solutions. The purpose of stage two is to provide a prediction of the minimum DST value as the geomagnetic storm develops as well as deliver a ranking of which predictor variables from the base-level are most important for prediction. To test this framework,

a dataset is created comprised from various data sources relevant to geomagnetic storm prediction. In addition, because many of these predictor variables have been well-studied, performance of the proposed variable importance ranking can be assessed. Therefore, the main idea of this work is to present a meta-learning framework that not only predicts with competitive accuracy against popular techniques, but delivers interpretation for the user on a representative, space weather dataset.

The subsequent sections of this work read as follows. Section 3.2 briefly reviews some previous studies on predicting geomagnetic storms as well as some background information beneficial for describing the methodology. Section 3.3 provides detailed explanations of how the experimental dataset is constructed, the proposed framework, and the implementation strategy. Section 3.4 displays results of predictive performance and variable importance scores. Sections 3.5 and 3.6 conclude with a summary and postulates areas for future work.

3.2 Literature Review

3.2.1 Forecasting Geomagnetic Storms

In general, a geomagnetic storm has three phases that can be seen in the data: a sudden commencement, a main phase, and a recovery [10]. The sudden commencement is derived from a abrupt uptick in solar wind dynamic pressure, which is a function of its speed and density, indicating the arrival of the shock from a CME [3]. The main phase begins when the interplanetary electric field (an interaction between the solar wind velocity and the southward magnetic field component) becomes large and positive [10]. This is a result of magnetic reconnection [22] [25] [29] (as cited in [90]). During the reconnection, particles are injected into Earth's magnetosphere. This, in turn, enhances Earth's ring current, or the high energy current within the magnetosphere [36]. The intensification of the ring current is indirectly proportional to the equatorial geomagnetic field [54]. Hence, the larger the enhancement of the ring current, the more shrinkage of Earth's protective magnetic field, leading to more significant geomagnetic storms [36]. Recovery is initiated when the injection rate decreases enough for the ring current to

start reverting back to normalcy. The process can be visualized by assessing the DST [79]. Expressed in nT and recorded every hour from observatories around the world, it measures the depression of the equatorial geomagnetic field, or horizontal component of the magnetic field; thus, the smaller the value of the DST, the greater the depression. [36].

While many works have focused on forecasting this index value (see [2] for a brief review of these), emphasis is primarily placed on using only IPI. Incorporating the initially observed CME characteristics can improve geomagnetic storm prediction [21]. The combination of this data has been incorporated in logistic regression models [74] [14], neural networks [85] [86], and random forests [47]. Further improvements can potentially be made by using multi-step frameworks. To narrow the scope, this work will focus on reviewing two recent multi-step procedures that predict geomagnetic storms using both near-Earth IPI and information taken near the Sun.

Valach et al. [86] reinforced one of the primary issues facing geomagnetic storm prediction: the inability to estimate the orientation of the interplanetary magnetic field from an incoming CME more than a few hours out. It is well-known that one of the largest predictor variables is the magnitude of the southward-directed magnetic field component B_z [10] [74] [36]; however, this is difficult to predict prior to reaching the L1 Lagrangian point due to complexities in a CME's magnetic topology [70]. Hence, under the assumption that the direction of the magnetic field component is unpredictable, the authors first studied the behavior of B_z for 2,882 days between 1997 to 2007 before implementing any predictive construct. Based on their analysis, they determined that for the majority of the days with a high-level of geomagnetic activity, B_z was negative for at least 16 hours during the course of the day (roughly 31% of the days studied). Then, after building a neural network using these observations, they forecast the daily level of geomagnetic activity with initial CME and solar X-ray information. The benefits to their approach are that the predictions are timely (absence of IPI in the second step enable forecasts at least a day out) and are well-suited for the strongest of storms (since the training observations are composed of days where B_z is negative for more than 16 hours). However, as noted by the authors, it does not do as well differentiating moderate and weak geomagnetic

storms. In addition, the time scale of the prediction is in days, which is not as granular as hours.

Kim et al. [40] argued that only using information based on urgent warning IPI for prediction does not provide a practical lead time, even though the forecasts are accurate. At the same time, strictly employing CME data becomes frivolous as each CME experiences changes in composition as they propagate through the interplanetary medium, thereby making prediction difficult. Therefore, Kim et al. [40] constructed a two-step forecasting system using both urgent warning IPI and initial CME data. At the first step, they applied multiple linear regression models to predict the strength of geomagnetic activity for northward and southward events at the onset of a CME using its location, speed, and direction parameter (estimated from the magnetic orientation angle of the related active region on the Sun). The estimation of the direction (north or south) is based on the assumption that these rarely deviate from that of the associated active region [95]. Next, they administered a set of rules based on the IPI to update the forecast and classify the impending CME as causing a moderate or intense storm. This method contributes a medium-term to short-term forecast from the first observance of a CME to its approach to Earth. While this method yields accurate and interpretable results, only 55 CMEs from 1997-2003 were studied. Moreover, the absence of using a validation scheme when constructing the rules can lead to overfitting when predicting on future data [33].

Interestingly enough, the former work assumes the direction of the magnetic field component in a CME is unpredictable while the latter does not for their step one models. In this work, the direction is not considered in any of the steps. Instead, the constructed dataset captures values of B_z prior to the climax of a given geomagnetic storm [47]. Thus, if this value is high in magnitude, then this reflects the southward behavior. More details about the dataset constructed are given in Section 3.3. Aside from the work by Dryer et al. [21], which used an ensemble of four physics based models to predict shock arrival times, the idea of using ensemble models has not been very prevalent in the literature. Therefore, leveraging more advanced ensemble frameworks for predictive modeling has

the opportunity to increase accuracy in this field.

3.2.2 Meta-learning via Stacked Generalization

As mentioned before, each predictive algorithm or model invokes a certain bias. The goal of stacked generalization is to exploit these various biases to create a more adaptive learning scheme [8]. The idea can be simplified in the following steps [92]:

- Construct a dataset consisting of predictions from a set of level 0 (or base) learners using a training and a test set. Refer to this as the metadata, *MD*.
- Generate a level 1 (or meta) learner that utilizes the predictions made at the previous level as inputs. That is, train the meta-learner on *MD* as opposed to the original training data.

The above process is visualized in Figure 3.1.

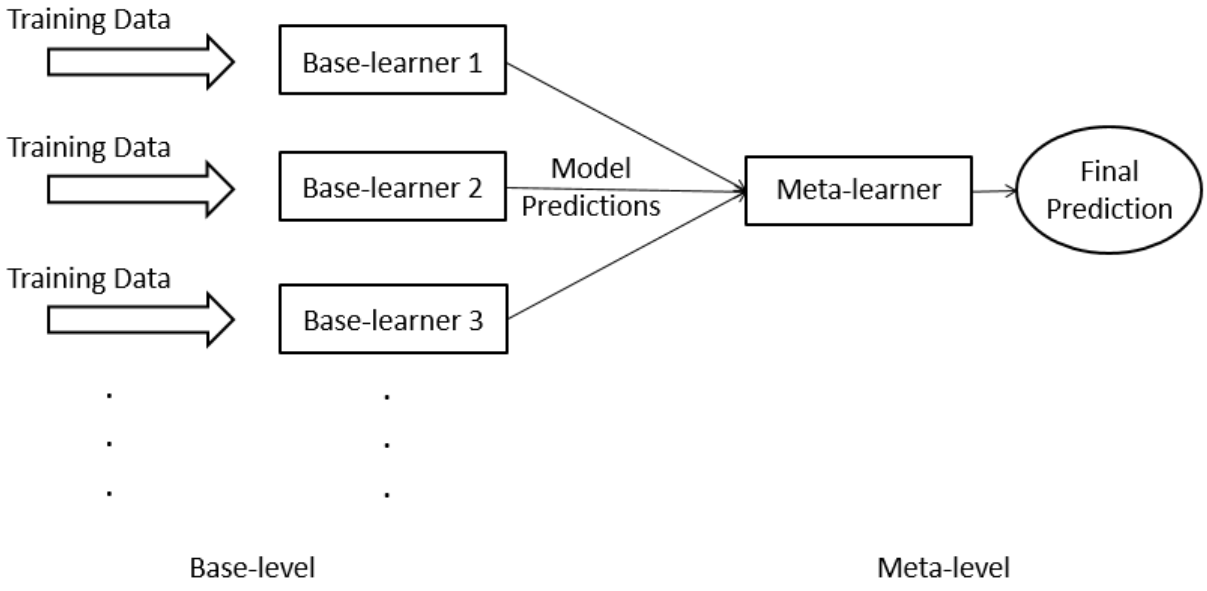


Figure 3.1: Diagram of stacked generalization.

Often times, the predictions from the base-learners are determined via k-fold cross-validation [9] [91]. Define the dataset $S = \{(y_i, \boldsymbol{x}_i), i = 1, ..., n\}$ where $\boldsymbol{x}_i$ is a vector of predictor variables and y_i is the corresponding response value for n observations. Specifically, split the dataset S into k near equal and disjoint sets such that $S_1, S_2, ..., S_k$. Let $S^{-k} = S - S_k$ and S_k to be the training and test sets, respectively. Execute the base-learner on the first S^{-k} parts to produce prediction for the held-out part S_k. Repeat this

procedure until each subset of S has been used as a test set exactly once. Extract all the hold-out predictions to create *MD*.

The meta-learner's purpose is to gain information about the generalization behavior of each learner trained at the initial level. This learning process can be enhanced when the base-level consists of a diverse selection of techniques, each generalizing the dataset in a different way. While theoretically inciting, Wolpert explains that this process is a "black art" since no rules are derived as to the selection of what models or algorithms to use at either level nor the input space for the top level generalizer. Ting and Witten [81] attempted to resolve via real data studies. Their conclusions showed that the best meta-learner to enforce for classification tasks is a multi-response linear regression algorithm with class probabilities (as opposed to class predictions) as its inputs. This model separates multi-class problems with R number of classes into R regression problems. The response for the regression model for class c is constructed via a binary split, where an observation is either class c or zero otherwise. Thus, the linear regression problem for class c is

$$LR_c(x) = \sum_{m}^{M} \beta_{mc}\pi_{mc}(x) \tag{3.1}$$

where β represents the coefficients calculated from a non-negative least square algorithm [48] and π symbolizes the class probability prediction of the c^{th} output class from the m^{th} base-learner. Then, after enumerating $LR_c(x)$ for all R classes, each new observations x can be appointed to class c if

$$LR_c(x) > LR_{c'}(x) \ \forall \ c' \neq c \tag{3.2}$$

Besides accuracy, the advantage to using this model is the ability for interpretation. That is, the user may glean the contribution of each base-learner by analyzing the magnitude of the coefficients. While the non-negativity constraint did not make much of a predictive difference for classification tasks, it has been shown to improve error in pure regression settings [49] [9].

Todorovski and Džeroski [82] introduced the concept of meta decision trees. Tradi-

tionally, a decision tree determines a class prediction in its leaves. For this model, the leaves establish which base-learner to utilize for prediction based on predicted class probability distribution properties such as the entropy and the maximal probability. This allows for the base-learner that has the best relative "area of expertise" on a subset of training data given in the leaves to be used for prediction.

Others have implemented different types of linear methods at the meta-level. Sill, Takacs, Mackey, and Lin [72] combined the predictions from different base-learners using a feature-weighted linear stacking algorithm. They were able to improve their results by incorporating a set of 25 meta-level predictor variables in conjunction with their model outputs to achieve a submission performing as well as the winners of the Netflix Prize competition. Sam Reid and Greg Grudic [65] explored the issue of overfitting within stacked generalization as noted by Caruana, Niculescu-Mizil, Crew, and Ksikes [13]. They experiment with ridge (L_2 penalty) [34],

$$P_{ridge} = \sum_{j=1}^{p} \beta_j^2 \leq t \tag{3.3}$$

lasso (L_1 penalty) [80],

$$P_{lasso} = \sum_{j=1}^{p} |\beta_j| \leq t \tag{3.4}$$

and elastic net (mixture of L_1 and L_2 penalties) [100]

$$P_{enet} = \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2) \tag{3.5}$$

where α is a convex combination defining the weights for each penatly on p predictor variables. These were instituted as meta-learners for around 1,000 base-learners using the StackingC method for several multi-class datasets. They showed that ridge regression performed the best on the majority of the datasets. The authors demonstrate the need for regularization to improve generalization accuracy, especially in the presence of highly correlated outputs and well-tuned models, for multi-class problems.

Implementing lasso and elastic net penalties offer the opportunity for variable selec-

tion at the meta-level. The concept of reducing the size of ensemble models has been researched in other works [96] [98] [67]. For instance, Rooney, Patterson, and Nugent [67] investigated approaches based on maximizing the accuracy and diversity of the base-learner predictions. While pruning the ensemble size yielded similar error compared to using all of the base-learner predictions on average, it did not match the un-pruned counterpart on every dataset. The authors noted this could be due to only utilizing one set of parameter values, instead of tuning to adapt to each dataset tested. This shows the importance of parameter tuning as well as the possibility of gaining more parsimonious and general solutions via regularization.

Recent uses of stacked generalization have been employed in the financial sector. Abbasi, Albrecht, Vance, and Hansen [1] developed a meta-learning framework for detecting financial fraud using publicly available data. The authors focused on reducing bias by preparing a diverse set of financial indicators and applying an adaptive stacked generalization algorithm. They selected 14 classifiers to serve at the base-level and the meta-level. In addition, they integrated error cost settings so as to place more weight on committing a Type II error (false negative) as opposed to Type I error (false positive) for regulators and investors. Their MetaFraud method outperformed all other fraud detection systems tested. Tsai and Hsu [83] discussed a scheme based on stacked generalization to predict bank failure. Unlike other works, the main purpose of the base-learners was to filter out the "noisy and unrepresentative training data for the level 1 classifier" (p. 171). Those observations with better predictive agreement served as inputs for the meta-learner. This data reduction technique helped the framework yield the highest accuracy rates in comparison to other single classifier methods and traditional stacked generalization.

Based on the results in previous studies and given the nature of the base-learner predictions, it seems advantageous to implement a regularized meta-learner to have the best potential for success in stacked generalization. By using various types of penalty functions, a learning system can effectively make predictions and provide sparse solutions, even in situations with high amounts of multicollinearity. However, none of the studies mentioned above discuss how to choose a meta-learner when the outlier values are im-

portant for regression tasks. Specifically for predicting geomagnetic storms, outliers are important because strongly geoeffective CMEs do not occur often; hence, a meta-learner cannot downplay the effect of these for prediction. If anything, the meta-learner should treat these values with more emphasis. In addition, subsetting the data to only include these outliers for model construction inhibits meta-knowledge to be gained for all CME events. Therefore, for this study, a regularized quantile regression model is chosen for the meta-learner in order to more adequately deal with outliers, improve accuracy, and promote sparse solutions.

3.2.3 Regularized Quantile Regression

Recall the ordinary least squares (OLS) solution for the coefficients in linear regression:

$$\hat{\beta}^{ols} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{3.6}$$

where $\boldsymbol{X}$ is the predictor matrix of dimension $n \times (p+1)$ and $\boldsymbol{Y}$ is the vector of outcomes of dimension $n \times 1$ for n observations and p predictors. Specifically,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Alternatively, Eq. 3.6 can be written as the following optimization problem:

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \tag{3.7}$$

It is often popular to apply some sort of regularization, or penalty, to the estimated coefficients such as in ridge, lasso, and elastic net regression. In general, this can be expressed as [61]

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{3.8}$$

where $p_\lambda(\cdot)$ dictates the type of penalty function with a non-negative constant λ to deter-

mine the amount of regularization. Utilizing constrained regression approaches enables the ability to perform variable selection or improve prediction in particular environments. However, the main goal in these methods is to estimate the conditional mean of some response given a set of predictor variables. Situations may arise where it is more advantageous to investigate a certain part of the conditional distribution [57] [11]; hence, quantile regression was developed [44]. The goal of quantile regression is to "offer a comprehensive strategy for completing the regression picture" (pg. 1) [43]. In general, this involves minimizing the sum of asymmetrically weighted absolute residuals [44]

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \quad \frac{1}{n}\Bigg[\sum_{i\in\{i:y_i\geq \boldsymbol{x}_i'\boldsymbol{\beta}\}} \tau|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}| + \sum_{i\in\{i:y_i< \boldsymbol{x}_i'\boldsymbol{\beta}\}} (1-\tau)|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}|\Bigg] \tag{3.9}$$

for some given quantile level τ. In this way, different weights are placed on positive (under-prediction) and negative (over-prediction) errors corresponding to the desired quantile. Note that when $\tau = 0.5$ this simply reduces to median regression. As with the linear case, the coefficients in quantile regression can be penalized the same way. Using lasso has been a popular choice due to its sparse nature [42] [50]. However, it has been shown that lasso has some limitations in high-dimensional situations or ones with severe multicollinearity [100]. In addition, it lacks oracle properties [26] [99]. That is, lasso does not select the correct subset of predictor variables while also efficiently estimating the non-zero coefficients as if only the truly influential predictor variables are included in the model, asymptotically [5]. Thus other penalties, such as adaptive lasso [99] and smoothly clipped absolute deviation (SCAD) [26] have been developed. These have both been shown to retain oracle properties for penalized quantile regression models [93]. The latter penalty can be defined as a quadratic spline function with knots at λ and $a\lambda$ to make the following objective function:

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \quad \frac{1}{n}\Bigg[\sum_{i\in\{i:y_i\geq \boldsymbol{x}_i'\boldsymbol{\beta}\}} \tau|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}| + \sum_{i\in\{i:y_i< \boldsymbol{x}_i'\boldsymbol{\beta}\}} (1-\tau)|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}|\Bigg] + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{3.10}$$

where

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & 0 \leq |\beta| < \lambda \\ \dfrac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} & \lambda \leq |\beta| \leq a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & |\beta| > a\lambda \end{cases}$$

for some $a > 2$ and $\lambda > 0$. By assigning different weights depending on $|\beta|$, SCAD avoids over-penalizing large coefficients, as is a common problem in lasso [26] [93] [61]. However, unlike the adaptive lasso, the SCAD penalty makes solving Eq. 3.10 more difficult due to its non-convex nature. Fortunately, efficient algorithms have been developed to increase the computational speed for solving these non-differentiable and non-convex optimization problems [61]. Because of the advantages of using the SCAD penalty, this work employs this type of regularization on a quantile regression model at the meta-level. Note that subsequent uses of "SCAD" refer to the quantile regression model in Eq. 3.10.

3.3 Methodology

3.3.1 Data Preparation

The goal of this data collection is to obtain a large number of CME events from 1996-2014. Four main sources of data are considered: near-Earth CME information provided by Richardson and Cane [12] [66] (referred to as "ICME list"), OMNI hourly averaged solar wind data at one AU from the Coordinated Data Analysis (Workshop) Web [41] (referred to as "OMNI database"), CME measurements given by the LASCO located on the SOHO satellite [30] (referred to as "LASCO catalog"), and daily Sun properties recorded by NOAA [59].

3.3.2 The LASCO Catalog and the Sun Properties

Data from the LASCO catalog [30] serves as the basis for collecting CMEs. All events between 1996-2014 are collected. Gopalswamy et al. [30] viewed the basic characteristics of a CME are their linear speed, angular width, central position angle, and acceleration.

Because the central position angle does not exist for full halo CMEs (where angular width is 360 °), the measurement position angle is used instead. Additionally, the three second-order speeds and comments are included. The comments are investigated in the dataset by creating two binary variables indicating whether a CME is considered "poor" or "very poor." These are mutually exclusive by design. Along with information from the LASCO catalog, daily solar weather data are obtained from NOAA [59]. This information includes the average solar radio flux, number of sunspots, sunspot area, number of new sunspot regions, and a count of the C, M, and X-class solar flares for that day. Since these data are daily aggregations, they are merged with the LASCO catalog by the CME expulsion day.

3.3.3 Predicting Arrival Time

In order to create a training dataset, a DST value must be assigned to each of these CME events. Assigning these values to a CME event is usually done manually like in the ICME list. Richardson and Cane [12] [66] report the disturbance arrival time, the CME magnetic field leading edge, and the respective trailing edge of the mass in their dataset. The authors defined the leading and trailing edge of the CME themselves from examining magnetic field and plasma data [12]. Using the disturbance arrival time and trailing edge, the associated strength of each CME is captured by recording the lowest DST value to occur in that interval. The resulting effort is a list of near-Earth CMEs, which is still being updated currently.[1] However, because it is composed of only near-Earth CME events, only a small proportion of all CME event occurrences are accounted for in this list. Therefore, for the purposes of this study, it is necessary to also consider CMEs which do not directly impact Earth so as to provide a more realistic prediction scenario. An inclusion of a larger number of CMEs than in previous studies renders manual imputation of the associated storm strength and the IPI quite infeasible. To overcome this difficulty, approximating the arrival time of an impending CME is vital.

Previous studies have implemented empirical models using CME acceleration [31], regressions with the CME initial speed [97] [90] [75], or more recently with neural networks

[1] Table can be found online at `http://www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm`

using initial speed and Central Meridian Distance [78]. Due to the difficulties in predicting arrival times, using regression with linear speed is chosen [75]. Define T to be the transit time and LS to be the linear speed found in the LASCO catalog. Previous models include

$$T_{hours} = 27.98 - \frac{2.11 \times 10^4}{LS} \tag{3.11}$$

from Wang, Ye, Wang, Zhou, and Wang [90] based on 15 CMEs from March 1997 to December 2000,

$$T_{hours} = 96 - \frac{LS}{21} \tag{3.12}$$

from Zhang, Dere, Howard, and Bothmer [97] based on 26 CMEs from 1996 to 2000, and

$$T_{hours} = 86.9 - 0.026 \times LS \tag{3.13}$$

from Srivastava and Venkatakrishnan [75] based on 64 CMEs from 1996 to 2002. The latter authors indicate that a simple linear relationship is most appropriate as opposed to the non-linear approach taken by Wang et al. [90]. Therefore, this work posits a new linear relationship between LS and T using more CME events.

In order to formulate a new equation, the transit times need to be calculated. This can be done by using the ICME list. In the ICME list, the probable CME association from the LASCO catalog is included for events where it can be detected. For those CME associations in which time discrepancies exist between the LASCO catalog and the ICME list, the closest CME is elected. For instance, according to the LASCO instrument, at 21:30:08 UT on February 17th, 2000, a halo CME was recorded. In the ICME list, the authors indicate a halo CME association at 20:06:00 UT on that same day. Since no other halo CME occurred on that day, it is possible that the official time of this CME in the LASCO catalog may have been updated since the authors received the data. Moreover, for instances in which LASCO detected multiple CMEs at the exact same time, the most likely CME is chosen based on the remarks in the LASCO catalog. For example, one may be commented as a "poor event," so consequently, the other is chosen as the probable

CME association (see August 6th, 2000, at 18:30:32 UT).

Once those CMEs in the ICME list are linked with those from the LASCO catalog, 201 events are used to construct the simple linear regression model. Transit times may be calculated by subtracting the LASCO CME detection time from the disturbance arrival time found in the ICME list. Regressing the linear speed on this newly created transit time yields the following equation:

$$T_{seconds} = 304555 - 90.57 \times LS \tag{3.14}$$

The negative slope is consistent with the other two linear regression models stated above. Note that larger coefficients are due to using seconds as opposed to hours. To emulate the CME interval of its leading and trailing edge found in the ICME list, the 95% prediction interval from Eq. 3.14 is defined. Once the model is built, it is then applied to all the CMEs in the LASCO catalog. This provides an upper and lower bound for arrival time for each CME event, regardless of whether they are near-Earth or not. While it may seem trivial to calculate the arrival time interval of a CME that does not impact Earth, this step is necessary in order to acquire the geomagnetic storm index and associated IPI as will be shown in the next section.

3.3.4 Assigning DST Index and IPI

To assign the geomagnetic strength index to a CME event, a similar strategy is implemented as in the ICME list. Starting at the lower bound of the prediction interval and ending at the upper prediction bound, the lowest value of the DST is recorded from the OMNI database. This is done automatically by programming a loop to search within each calculated prediction interval for each CME event obtained from the LASCO catalog, regardless of their proximity to Earth. Note that if a CME does not interact with Earth, then its associated DST value will likely not indicate much geomagnetic disturbance.

A similar process is used for the related IPI. Kim et al. [40] instituted the OMNI database in their study to obtain their IPI. They find the lowest value of B_z within their time interval as well as the lowest DST value. B_z typically minimizes prior to the

minimization of the DST value [55] [40], which is why it is an extremely useful predictor variable. They also record the lowest B_z value starting from 24 hours prior to their expected CME arrival time and ending at the time of the minimum DST value. Similarly in this work, another loop is constructed to obtain the minimum B_z value starting from the lower bound of prediction interval from Eq. 3.14 and ending at the time of the minimum DST value found by the first loop. Along with B_z, other relevant IPI used in other studies [40] [2] are included at the time of the B_z minimum.

While this automatic system makes data collection easier, two caveats exits:

1. Low values of B_z may be recorded, even if the respective CME did not approach Earth. Such disturbances can come from other from solar sources such as co-rotating interaction regions [97]

2. For multiple CMEs whose time intervals overlap, it is possible that a weak CME has been given the same strong IPI values of a known strong CME. It is also possible that CMEs may share identical values.

Naturally, the best way to assign these IPI values to a CME is by manually looking at the relevant information. However, since the goal of this work is to study a larger number of CMEs, parsing these out is infeasible. To help combat the effect of the second caveat, non-near-Earth CME observations with duplicated IPI extraction times are omitted. This greatly reduces the noise surrounding strongly geoeffective CME events and their IPI values as well as in the those that do not approach. While again it may seem trivial to be concerned with the values of non-near-Earth CMEs (since they should not be of significance), these values are necessary to produce a quality dataset. Hence, only including CMEs with distinctly defined IPI by the automated process will enable cleaner and clearer dataset to test the proposed meta-learning framework.

Moreover, CME events in which the automated process recorded a DST ≤ -200nT but are not classified as near-Earth via the ICME list are also eliminated. This is likely due to their values being obtained within the predicted time interval from a near-Earth CME that launched shortly after or before the CME in question. Therefore, these are also discarded since severe geomagnetic storms are typically induced by CMEs [32] [6] [84] [97].

Furthermore, near-Earth CMEs that share identical IPI values are inspected against the ICME list to distinguish which CMEs are truly severe. Thus, if a CME shares a similar DST value to that recorded in the ICME list, it is included in the analysis while the other CME sharing its assigned IPI values is rejected. This prevents affirming that a CME is strongly geoeffective when in actuality it is not. Once all the predictor variables are collected, any rows with missing values are excluded to avoid ambiguity [40]. Thus, a master dataset is realized composed of 2,811 CME events. Approximately 5% of these are deemed as strongly geoeffective. The list of predictor variables is divided into initial CME and solar characteristics (Table 3.1) and the subsequent IPI (Table 3.2).

Two exceptions exist for omitting events with missing data: October 28th 11:30:05 and October 29th 20:54:05 CMEs (see LASCO catalog). These were major events in the infamous "Halloween storms" in 2003. These caused significant black-outs in Sweden leaving tens of thousands without power [63]. However, satellite instruments were unable to capture explicit values for some of their important IPI. In an effort to make sure these are included in the analysis (since each yielded a DST of -353nT and -383nT, respectively), V_{sw} is substituted with the average CME speed from the ICME list for these two events. This substitution allows for E_y (the interaction of ${B_z}^2$ and V_{sw} denoted by $E_y = [-V_{sw} * B_z]^{10^{-3}}$) to be calculated. For the other missing IPI values, they are imputed with the average from all the other CMEs which produced a DST value ≤ -300nT. These two storms will be revisited in the results section to investigate how well the proposed meta-learning framework performs in assessing these events. Note that since this is a two-stage process, two datasets exist corresponding to both stages. Stage one encompasses classifying the CME at launch while stage two incorporates the important IPI to predict the minimum DST value before the climax of the geomagnetic storm. For the classification task, those with a DST of $<= -100$nT are labeled as "strong" while those > -50nT are considered "weak" and "moderate" otherwise [47]. A sample from the respective datasets may be found in Tables 3.3 and 3.4 representing CMEs from October and November of 2003. These two months represent a diverse collection of varying degrees

[2]Measured by the Geocentric Solar Magnetospheric Cartesian coordinates

x	Type	Description
MPA	C	Measurement position angle of CME at the height-time measurements (degrees)
AW	C	Sky-plane width of CME (degrees)
LS	C	Linear speed of CME (km/s)
SOI	C	Quadratic speed of CME at initial height measurement (km/s)
SOF	C	Quadratic speed of CME at final height measurement (km/s)
SOR	C	Quadratic speed of CME at height of 20 solar radii (km/s)
Acc	C	Acceleration of CME in (m/s^2)
$Poor$	B	Noted as a poor event in the comments
$Very\_Poor$	B	Noted as a very poor event in the comments
$RFlux$	C	Daily average 10.7cm flux values of solar radio emissions on CME ejection day in 10^{-22}J/s/m^2/Hz
SSN	D	Number of sunspots recorded on CME ejection day
SSA	C	Sum of the corrected area of all observed sunspots on CME ejection day in millionths of the solar hemisphere
NR	D	Number of new sunspot regions on CME ejection day
$XrayC$	D	Number of C-class solar flares on CME ejection day
$XrayM$	D	Number of M-class solar flares on CME ejection day
$XrayX$	D	Number of X-class solar flares on CME ejection day

Table 3.1: List of CME and Sun characteristics. Predictor variable types are indicated as C-Continuous, D-Discrete, B-Binary.

x	Type	Description
E_y	C	Interplanetary electric field in millivolts per meter (mV/m)
B_x	C	X-component magnetic field component (nT)
B_y	C	Y-component magnetic field component (nT)
B_z	C	Southward magnetic field component (nT)
V_{sw}	C	Plasma flow speed (km/s)
Phi	C	Plasma flow direction longitude (degrees)
$Theta$	C	Plasma flow direction latitude (degrees)
D_p	C	Proton density in Newtons per cubic centimeter (N/cm^3)
Na_Np	C	Alpha to proton ratio
T_p	C	Proton temperature in degrees Kelvin (K)
P	C	Flow pressure in nanopascals (nPa)
$Beta$	C	Plasma beta

Table 3.2: List of IPI. Predictor variable types are indicated as C-Continuous, D-Discrete, B-Binary.

of CME geoeffectiveness. The total list spans from January of 1996 through December 2014.[3]

3.3.5 Variable Importance for Stacked Generalization

As addressed by Wolpert [92] and Ting and Witten [81], parts of stacked generalization are a "black art." Consequently, the latter authors posited that using a linear model as a meta-learner is advantageous because it not only yielded the best results, but also delivered the ability for some interpretation by analyzing the weights placed on each base-leaner. This work seeks to go a step further to establish some knowledge about the predictor variables invoked at the base-level as well. This may be accomplished by extracting variable importance scores from each base-learner, provided they have a built-in variable importance scheme. This process will allow for insight into how useful each predictor variable is to each base-learner's own prediction.

Denote $\boldsymbol{b} = \{b_1, b_2, ..., b_m\}$ to be the set of m base-learners and $\hat{\boldsymbol{y}} = \{\hat{y}_{b_1}, \hat{y}_{b_2}, ..., \hat{y}_{b_m}\}$ to be the predicted DST value from each of the base-learners. Let $\boldsymbol{I}$ symbolize the $p \times m$ matrix of min-max normalized[4] variable importance scores corresponding to each respective base-learner on a scale from zero to 100. It is necessary to normalize such scores because of the varying methodologies of how different models and algorithms compute variable importance. Therefore, placing each base-learner's variable importance scores on the same scale allows for fair comparisons. Finally, note $\boldsymbol{\beta} = \{|\hat{\beta}_{b_1}|, |\hat{\beta}_{b_2}|, ..., |\hat{\beta}_{b_m}|\}$ to represent the vector of estimated coefficients from SCAD. Note that the absolute value is taken to represent the overall contribution in magnitude, regardless of whether it is positive or negative. The proposed strategy can be composed in the following way:

1. Obtain $\boldsymbol{I}$ from $\boldsymbol{b}$ and $\boldsymbol{\beta}$ from the SCAD.

2. Multiply $\boldsymbol{I}$ by $\boldsymbol{\beta}$ so that the score for each of the predictor variables is multiplied by the absolute value of the estimated coefficient assigned to each of the respective base-learners via SCAD. Hence, for $j = 1, ..., m$, each base-learner's variable

[3]Some CMEs after December 26th, 2014, are omitted since the upper prediction limit from Eq. 3.14 begins to fall into 2015 and the data from the OMNI database collected for this work spans only through 2014.

[4]$(X - X_{min})/(X_{max} - X_{min})$

LASCO Date:Time	DST	MPA	AW	LS	SOI	SOF	SOR	Acc	$Poor$	$Very_Poor$	$RFlux$	SSN	SSA	NR	$XrayC$	$XrayM$	$XrayX$
09OCT03:19:33:07	weak	289	6	507	592	425	0	-26.1	0	0	111	68	280	1	1	0	0
13OCT03:03:54:05	moderate	62	49	362	348	377	448	3.3	0	0	94	25	40	0	0	0	0
13OCT03:18:30:06	moderate	286	109	349	319	379	421	3.2	0	0	94	25	40	0	0	0	0
15OCT03:00:30:05	moderate	87	6	176	156	195	461	7.8	0	0	96	29	40	1	0	0	0
22OCT03:08:30:32	weak	352	267	719	403	1080	1009	37	0	0	154	117	1950	1	0	7	0
28OCT03:11:30:05	strong	15	360	2459	2686	2229	2268	-105.2	0	0	274	230	4520	1	5	0	1
29OCT03:20:54:05	strong	190	360	2029	2406	1670	1519	-146.5	0	0	279	330	5160	0	4	2	1
03NOV03:19:31:43	weak	109	26	641	560	722	948	25.3	0	0	167	76	2830	0	3	1	2
04NOV03:12:06:06	weak	84	360	1208	1467	926	1090	-41.2	0	0	168	79	1100	1	3	3	1
04NOV03:12:54:05	weak	263	72	605	185	1027	1038	44	0	0	168	79	1100	1	3	3	1
09NOV03:06:30:05	moderate	112	360	2008	2420	1579	1788	-128.6	0	0	93	47	100	1	0	0	0
12NOV03:18:30:05	moderate	249	88	891	998	778	697	-21.9	0	0	99	39	220	0	4	0	0
13NOV03:22:30:05	weak	142	113	554	654	457	304	-14.1	0	0	102	25	450	1	2	2	0
14NOV03:10:54:05	weak	281	57	683	817	535	0	-28.1	0	0	99	34	350	1	4	0	0
15NOV03:17:50:05	weak	267	148	1375	1276	1465	1513	28.1	0	0	98	52	360	1	3	0	0
16NOV03:07:50:05	weak	251	18	979	1032	930	792	-18.8	0	0	104	54	390	1	6	0	0
18NOV03:08:50:05	strong	206	360	1660	1674	1645	1656	-3.3	0	0	144	90	900	2	4	4	0
19NOV03:09:26:05	strong	334	84	422	261	581	628	13.7	0	0	155	114	2140	1	8	1	0
19NOV03:15:06:05	strong	209	48	606	409	806	1164	49.8	0	0	155	114	2140	1	8	1	0
20NOV03:08:06:05	strong	243	47	890	991	797	0	-44.7	0	0	175	118	2010	0	9	3	0
21NOV03:17:50:05	moderate	290	15	554	607	508	423	-8.1	0	0	177	131	1700	0	5	0	0
21NOV03:19:27:16	moderate	107	82	737	905	568	147	-32.9	0	0	177	131	1700	0	5	0	0
22NOV03:01:50:05	weak	246	18	675	801	549	0	-80.6	0	0	176	123	1660	0	6	0	0
23NOV03:09:18:05	weak	357	55	283	0	496	557	14.8	0	1	178	158	1570	0	7	0	0
24NOV03:10:50:06	weak	55	18	440	489	388	202	-8.2	0	0	177	149	1520	2	6	0	0
25NOV03:04:06:05	weak	75	43	791	789	793	794	0.3	0	0	171	202	1320	2	4	0	0
30NOV03:12:26:05	weak	282	13	495	110	864	1615	109.3	1	0	153	178	1070	1	4	0	0

Table 3.3: Data sample for the training data in stage one of the two-stage procedure. Note that since stage one is a classification problem, DST represents a category of index strength, not the actual value.

LASCO Date:Time	DST	E_y	B_x	B_y	B_z	V_{sw}	Phi	$Theta$	D_p	Na_Np	T_p	P	$Beta$
09OCT03:19:33:07	-31	2.05	-5.6	2.8	-5.5	372	-2	1.8	13.4	0.036	86253	3.54	1.27
13OCT03:03:54:05	-77	2.92	-4.3	3.7	-4.2	695	1.4	2.1	3.1	0.036	324321	2.86	0.75
13OCT03:18:30:06	-60	1.78	-5	2.5	-2.7	659	1.5	0.7	2.8	0.048	175951	2.42	0.74
15OCT03:00:30:05	-53	3.43	0	1.2	-5.9	581	-1.8	3.8	4.2	0.036	102758	2.71	0.99
22OCT03:08:30:32	-49	3.34	-3.6	-4.1	-7.4	451	-2.6	1.8	5.1	0.124	38514	2.59	0.39
28OCT03:11:30:05	-353	34.45	-28.5	6.4	-26.5	1300	0.66	0.22	11.24	0.0946	95818.4	11.664	0.046
29OCT03:20:54:05	-383	21.68	-19.4	-8.4	-27.1	800	0.66	0.22	11.24	0.0946	95818.4	11.664	0.046
03NOV03:19:31:43	-29	-2.06	-1	-4.4	3.6	571	-2.6	2.9	3.5	0.027	169617	2.11	1.21
04NOV03:12:06:06	-28	-2.13	0.1	-3.8	3.9	547	-1.4	1.3	3.2	0.028	139892	1.78	1.1
04NOV03:12:54:05	-25	0.82	4	-1.8	-1.7	484	-0.7	-3	3.2	0.026	68875	1.38	1.05
09NOV03:06:30:05	-62	4.8	1.6	-0.1	-8.6	558	-0.7	1.8	6.8	0.022	149857	3.85	0.87
12NOV03:18:30:05	-55	3.05	-4.6	1.3	-4.8	636	3	3.4	4.6	0.036	205802	3.55	0.72
13NOV03:22:30:05	-49	2.27	-5.7	5.7	-3.5	649	-2.3	0.5	3.4	0.031	252280	2.69	0.6
14NOV03:10:54:05	-49	2.6	-6.4	1.3	-3.8	685	-0.4	2	3	0.017	190021	2.51	0.52
15NOV03:17:50:05	-48	2.83	-4.1	3	-3.7	764	-0.2	1	1.8	0.014	318093	1.85	0.82
16NOV03:07:50:05	-42	2.57	-3	1.3	-3.4	757	0.4	1.1	2.1	0.017	302207	2.15	0.98
18NOV03:08:50:05	-422	31.25	0.9	22.5	-50.9	614	1.3	-1.3	5	0.061	64842	3.92	0.01
19NOV03:09:26:05	-191	0.73	-4.9	10.4	-1.3	562	3.3	-0.9	8.8	0.117	80394	6.81	0.51
19NOV03:15:06:05	-185	0.5	-7	6.8	-0.9	554	3.5	-0.4	9.8	0.061	95280	6.25	0.84
20NOV03:08:06:05	-105	2.12	6.8	-2.8	-4.2	504	-3.2	-2.4	0.9	0.03	70711	0.43	0.1
21NOV03:17:50:05	-66	0.48	7.9	-1.4	-0.9	538	5.5	-1.8	4	0.042	229214	2.26	0.66
21NOV03:19:27:16	-71	2.22	6.1	-4.3	-4	554	2.5	-1.6	3.8	0.034	222517	2.21	0.58
22NOV03:01:50:05	-47	1.27	6.4	-4.2	-2.3	554	0.5	-2.1	3.7	0.043	164148	2.22	0.52
23NOV03:09:18:05	-31	0.94	4.8	0.6	-1.6	588	3.4	-1.1	3.7	0.061	167409	2.66	1.18
24NOV03:10:50:06	-24	0.15	5.2	-0.2	-0.3	502	2.2	0.8	3.4	0.028	178379	1.59	1.49
25NOV03:04:06:05	-22	0.8	5.1	0.7	-1.6	497	4.6	-0.3	2.7	0.03	161312	1.25	1
30NOV03:12:26:05	-17	0.68	3.2	-4.4	-1.6	423	-2	-1.7	4.4	0.054	114868	1.6	1.05

Table 3.4: Data sample when the important solar wind parameters are added. Note that the dataset for stage two includes that from stage one plus these. Further note that the response is now the DST value and not the classification as found in Table 3.3.

importance scores will either be

- increased if $|\hat{\beta}_{b_j}| > 1$
- decreased or stay the same if $0 < |\hat{\beta}_{b_j}| \leq 1$
- reduced to zero if $|\hat{\beta}_{b_j}| = 0$

	b_1	b_2	b_3	...
p_1	100	85	90	...
p_2	80	65	75	...
p_3	50	55	80	...
⋮	⋮	⋮	⋮	⋱

Example of the data from $\boldsymbol{I}$.

	$\boldsymbol{\beta}$
b_1	2.53
b_2	0.77
b_3	0.00
⋮	⋮

Example of the data from $\boldsymbol{\beta}$.

	$\boldsymbol{I} \cdot \boldsymbol{\beta}$					
p_1	100(2.53)	+	85(0.77)	+	90(0.00)	...
p_2	80(2.53)	+	65(0.77)	+	75(0.00)	...
p_3	50(2.53)	+	55(0.77)	+	80(0.00)	...
⋮			⋮			⋱

Example of calculation from step 2.

3. Min-max normalize the scores from step 2. This creates a final variable importance score, I_{final}, ranging from zero to 100 for each predictor variable in the training data of dimension $p \times 1$.

After these steps are implemented, the product is a list of scores dictating how influential each predictor variable is in constructing $\hat{\boldsymbol{y}}$ conditioned on the utility of those base-learners at the meta-level. In devising an approach to extract knowledge at the base-level, less ambiguity shrouding the role of these predictor variables is realized when using stacked generalization.

3.3.6 Two-stage Meta-learning Framework

Figure 3.2 depicts the proposed two-stage meta-learning framework to predict geomagnetic storms. In the first stage, the dataset used (denoted as D_1) is comprised from three data sources: the LASCO catalog (for initial CME data), the ICME list (used to help

approximate CME arrival times and clean the training data), and daily Sun information from NOAA.

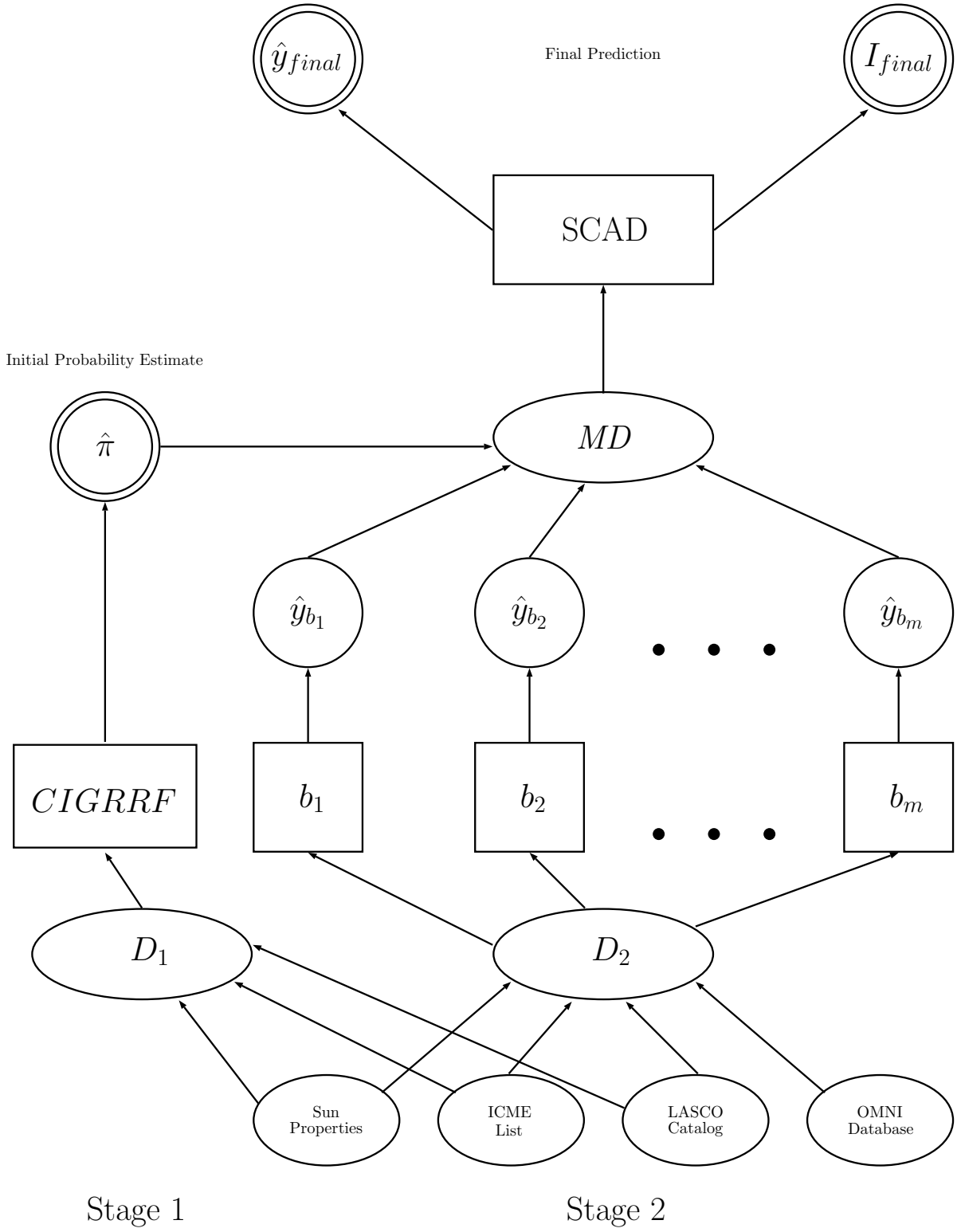


Figure 3.2: Diagram of the two-stage meta-learning framework for predicting geomagnetic storms.

The goal is to classify if a CME will produce a strong, moderate, or weak geomagnetic

storm at its launch. Recently, the idea of using CIGRRF to classify CMEs has been shown to be effective in terms of accuracy and interpretability [47]. Thus, this model is used here to estimate the probabilities of geoeffectiveness as well as tabulate some of the properties (maximal probability and entropy) about the distribution. This information can be used to make some preliminarily preparations days out as the CME propagates towards Earth (e.g. if the probability of producing a strong geomagnetic storm is high). In the second stage, the dataset used (denoted as D_2) contains all the information from D_1 plus the essential IPI taken when the CME is much closer to Earth. From here, the m base-learners can be executed to obtain forecasts of DST via cross-validation. Note that since this is an independent process, this can be done in parallel to increase computational efficiency. Once all of the predictions are collected, these can be combined with the probabilities and their properties from the CIGRRF to create *MD*.[5] Including this additional information can help make better predictions at the meta-level. Finally, SCAD is executed on *MD* to provide the final prediction and variable importance scores.

3.3.7 Improving on Biases

As noted before, stacked generalization seeks to build a more intelligent learning system that exploits the biases of the base-learners through use of a combining strategy. Two main sources of bias exists: declarative and procedural. Declarative involves the representation of the hypothesis space (i.e. the shrinkage or expansion of the feature space). Procedural relates to finding a suitable model or collection of models for a prediction task. In other words, declarative bias is driven by the types and number of predictor variables utilized for a task while procedural bias emphasizes how to best combine the base-learner predictions. Proper meta-learning systems effectively allocate the correct amount of bias for a particular problem [89] [8] [1]. Abbasi et al. [1] improved on these biases by incorporating a rich set of predictor variables at the base-level as well as employing stacked generalization with an adaptive learning algorithm to predict financial fraud.

In this work, declarative bias is managed in two ways: combining multiple data sources

[5]That is, the estimated probabilities for producing weak, moderate, and strong geomagnetic storms; the entropy of the three probabilities estimates; and the probability with the highest value are added to *MD*.

for base-level training and combining the stage one outputs alongside the model predictions in stage two. That is, the feature space is expanded at the base-level as well as at the meta-level to allow both the base-learners and the meta-learner to have plenty of information. Furthermore, procedural bias is improved by using stacked generalization to find the best mixture of base-learners to optimize the error rate. As acknowledged earlier, a weakness of stacked generalization is that fixed-form biases are exhibited by both the base-learners and the meta-learner. However, by shrinking the coefficients via regularization and selecting which base-learners to implement in a given problem, dynamic bias selection is taking place for the final prediction and variable importance scores by changing the representation of the feature space [89], which is at the core in constructing powerful meta-learning models. By improving these biases, we can hope to achieve some adaptability and, thus, better generalization error.

3.3.8 Implementation

Experiments with the proposed two-stage meta-learning framework are performed in R [64] version 3.3.3, mainly through use of the **caret** (Classification And REgression Training) package [28] [46]. This package allows for a streamlined user interface for applying a diverse set of predictive models from different packages with options to perform various pre-processing, post-processing, resampling, and visualization techniques. In addition, for those models that can perform variable importance estimation, the **caret** package can automatically extract these measures for a practitioner's use. Due to the large number of models available, a rich series of machine learning algorithms and statistical learning methods may be realized to construct the foundation of base-learners. Care is taken to ensure a diverse collection of 50 models or algorithms is used [96]. Unfortunately, in an effort to include a larger number of base-learners, not every model is able to provide model-specific variable importance scores. That is, they either do not have a way to calculate variable importance or **caret** does not implement one. For this study, only half of the base-learners have model-specific variable importance scores. For those that do not, the R^2 statistic is calculated using a loess smoother that is fit between the outcome and each predictor variable, as done by default within the package [46]. A

summary of the 50 chosen base-learners is listed in Table 3.5.

Model/Algorithm/Learner	**caret** Method	Model/Algorithm/Learner	**caret** Method
Bagged Multivariate Adaptive Regression Splines*	*bagEarthGCV*	Multivariate Adaptive Regression Splines*	*earth*
Bagged Regression Trees*	*treebag*	Neural Network*	*nnet*
Bayesian Additive Regression Trees*	*bartMachine*	Neural Network with Feature Extraction	*pcaNNet*
Bayesian Lasso	*blasso*	Non-negative Least Squares*	*nnls*
Bayesian Regularized Neural Network	*brnn*	One Standard Error Rule Regression Tree*	*rpart1SE*
Bayesian Ridge Regression	*bridge*	Partitioning Using Deletion, Substitution, and Addition Moves*	*partDSA*
Boosted Linear Model*	*glmboost*	Principal Component Regression	*pcr*
Boosted Tree	*bstTree*	Projection Pursuit Regression	*ppr*
Conditional Inference Random Forest*	*cforest*	Quantile Random Forest	*qrf*
Conditional Inference Tree	*ctree*	Quantile Regression with Lasso Penalty	*rqlasso*
Cubist*	*cubist*	Random Forest*	*ranger*
Extreme Gradient Boosting with Linear Booster*	*xgbLinear*	Regression Tree*	*rpart*
Extreme Gradient Boosting with Tree Booster*	*xgbTree*	Regularized Random Forest*	*RRFglobal*
Extreme Learning Machine	*elm*	Relaxed Lasso	*relaxo*
Generalized Additive Model using Loess*	*gamLoess*	Ridge Regression with Variable Selection	*foba*
Generalized Additive Model using Splines*	*gamSpline*	Robust Linear Model	*rlm*
Independent Component Regression	*icr*	SCAD Penalized Quantile Regression*	custom
k-Nearest Neighbors Regression	*knn*	Spike and Slab Regression	*spikeslab*
Kernel Partial Least Squares*	*kernelpls*	Stacked AutoEncoder Deep Neural Network	*dnn*
Lasso and Elastic Net*	*glmnet*	Stochastic Gradient Boosting*	*gbm*
Linear Regression*	*lm*	Supervised Principal Component Analysis	*superpc*
Linear Regression with Stepwise Selection	*leapSeq*	Support Vector Machines with Linear Kernel	*svmLinear*
Model Tree	*M5*	Support Vector Machines with Polynomial Kernel	*svmPoly*
Multi-layer Perceptron*	*mlpSGD*	Support Vector Machines with Radial Basis Function Kernel	*svmRadialSigma*
Multi-step Adaptive Elastic Net Regression*	*msaenet*	Weighted k-Nearest Neighbors Regression	*kknn*

Table 3.5: List of base-learners. Asterisk "*" denotes methods that have model-specific variable importance schemes via **caret**. Methods noted as "custom" signify that a custom model is created to introduce the learner into the **caret** framework.

Another advantage to using **caret** is the option to easily tune the parameters for a given learner by simply specifying a number for *tuneLength* in the *train* function. Each model has a predefined range of tuning values to search over proportional the *tuneLength*. The higher the *tuneLength*, the more tuning executed. The number of tuning parameters range for each model. In this experimental set-up, *tuneLength* is left at the default value of three.[6]

To implement the stage one CIGRRF, a custom model is created within **caret** using the *RRF* function in the **RRF** package [19] [20] [18] and the *cforest* function in the **party** package [35] [77] [76]. The parameter tuning approach used by Larkin [47] is adopted for this model with the best combination chosen by that which delivers the lowest LogLoss [68]. The only difference is in the final execution of the conditional inference random forest, where the number of trees is increased from the default 500 to 1000 in order to help stabilize variable importance estimates at stage one.

For the SCAD implementation, the **rqPen** R package is chosen [71]. This package

[6]The only exception to this is for the *bartMachine* method. Due to the large number of tuning parameter combinations executed at the default *tuneLength* (over 80), this is reduced to two.

offers estimation for SCAD as well as other penalized quantile regression models including lasso. In addition, it can utilize the recently proposed and efficient iterative coordinate descent algorithm [61] to compute SCAD solutions using the *QICD* function. Because this function is not offered in **caret**, it is incorporated within the **caret** framework by creating a custom model. Note that this custom model is also executed as a base-learner (see Table 3.5). To tune SCAD, only two parameters[7] are adjusted: the regularization value λ and the quantile level τ. The value of λ controls how much to penalize the coefficients and works similarly as λ in the popular **glmnet** package [27]. A wide range of values are investigated: $\lambda = \{1000, 1, 0.001\}$. For many applications of quantile regression, the selection of the quantile level τ is determined by the user to best fulfill his or her research goal. In this work, τ is treated as a tuning parameter to best find a balanced between accurately predicting the much rarer dangerous geomagnetic storms and much more common weaker counterpart. Quantile levels $\tau = \{0.1, 0.2, 0.3\}$ are selected since the 20^{th} percentile of the DST in the dataset is -49nT, which is approximately the threshold (-50nT) between weak and moderate storms for this work and in others [51]. Since the default amount of tuning is instituted, nine different parameter combinations for SCAD are tested. To benchmark the performance of using SCAD as the meta-learner, linear regression, non-negative least squares, and lasso/elastic net regression are also executed by calling the **caret** methods *lm*, *nnls*, and *glmnet* at the meta-level, respectively. These are selected as comparison models based on previous research in stacked generalization.

3.3.9 Resampling and Performance Metrics

As with any prediction problem, it is imperative to estimate the error rate on an independently held-out sample of data. Using the aforementioned k-fold cross-validation, this can be done by taking the average of the error from each of the held-out folds. It has been shown that using repeated k-fold cross-validation performs well in estimating the true error rate compared to other approaches such as .632+ bootstrapping [39]. However, when parameter tuning is involved, reporting the cross-validated estimate of error at a model's best parameter settings can be too optimistic; therefore, to carefully compare the

[7]The parameter a in Eq. 3.10 is left at the suggested default value of 3.7.

error rates, this work uses the nested cross-validation [88]. This procedure has two parts: an outer loop and inner loop of k-fold cross-validation. The outer folds represent the initial split of the data $S_1, S_2, ..., S_k$. The inner folds represents an internal k-fold cross-validation execution with training and tests folds constructed from S^{-k}. The purpose of the inner loop is to find the best parameters while the outer loop is to estimate predictive performance. Note that it is entirely possible to have different parameters chosen when evaluating the outer loop folds depending on the data partition [62]. For a comprehensive assessment of error, five repeats of 10-fold (5×10) nested cross-validation are conducted to best determine the performance of each model. To generate the data for *MD*, the hold-out predictions at stages one and two are extracted using just one iteration of the 10-fold nested cross-validation process. A final execution of the *train* function is conducted using all of the training data after the nested cross-validation procedures. Within this function, a final 10-fold cross-validation is performed to optimize parameters on all the data. This is to enable the variable importance scores to be extracted from CIGRRF, SCAD, and the base-learners based on all the data.

For many of the previous studies in predicting geomagnetic storms, the main performance metric utilized has been unweighted error criterion (e.g. root mean square error (RMSE) [37]). While RMSE does penalize larger errors more via the squaring operator, it treats each observation the same. In the context of predicting geomagnetic storms, it is more important for a model to accurately forecast the DST value associated for the strongest of storms. At the same time, focusing strictly on these observations can severely bias a model. Hence, the central metric used in this work for comparison and optimizing parameters is weighted mean absolute error (WMAE):

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i| \tag{3.15}$$

This is implemented within **caret** by creating a custom metric. Adopting WMAE allows for the opportunity to penalize models for inaccuracies when forecasting the more important observations (i.e., those CMEs that produce a geomagnetic storm with a DST ≤ -100nT). Considering only 5% of the observations in the constructed dataset meet

this precedent, it is necessary to implement. Given the potential impact that dangerous storms can have, strong CMEs are weighted as 10 times more important than the others (DST > -100nT). Using this 10:1 ratio is a conservative balance since strong geomangetic storms can result in economic losses in trillions of U.S. dollars [4]. In addition to WMAE, the overall RMSE and RMSE for the strong CME events will also be reported.

Given the added computational needs to perform stacked generalization, it is vital to investigate if simply using the best performing base-learner yields a similar error rate [23]. Therefore, careful methodology must be applied to inspect for statistical significance between the results. To accomplish this, the corrected repeated k-fold cross-validation test [7], or also known as the corrected t-test for repeated cross-validation [68], on the population of performance metrics (50 estimates from the 5×10 nested cross-validation) is used. This test has been shown to produce acceptable Type I error, low Type II error, and good replicability for comparing two models [7] [68].

3.4 Results

3.4.1 Predictive Performance

Table 3.6 denotes the errors metrics for the meta-learners as well as the base-learners. For space purposes, only the top 10 base-learners, according to WMAE, are shown.[8] SCAD yields the lowest WMAE compared to all methods, including the meta-learners and its execution at the base-level, with statistical difference when analyzing the RMSE on strong CME events. However, it only performs as the 6^{th} most accurate when looking at the RMSE for all the CME events and is out-matched with statistical significance by all three comparison meta-learners. It is important to reinforce that analyzing the overall RMSE alone can be misleading in this context. For instance, the best performing base-learner in stage one, cubist, achieves a lower error (RMSE = 18.21) compared to SCAD (RMSE = 18.68). However, looking at the strong CME RMSE reveals a much higher discrepancy (47.29 compared to 39.64). This occurs for the other meta-learners as well.

[8]SCAD outperforms the other 40 base-learners with statistical significance in all error metrics (with the exception of the quantile random forest); thus, it is not necessary to display these.

Hence, if a practitioner only considered this metric, a degradation in accuracy for the strong events will be realized. Therefore, for practical purposes in predicting geomagnetic storms, it is more appropriate to analyze error metrics such as the WMAE or subsets of RMSE, given the most costly and dangerous storms do not occur very often.

When considering the RMSE for dangerous CME events, SCAD reigns supreme. The only learner in which it does not significantly outperform for this metric is the quantile random forest. This makes sense since this model infers the full conditional distribution of a response variable [53]. Not surprisingly the most accurate base-learners are advanced ensemble models. In addition, it is not shocking to see improvements when using the regularized or constrained regression approaches compared to the unconstrained counterpart at the meta-level. Each of these obtained a better prediction by selecting only a certain number of base-learner predictions. On average, SCAD, non-negative least squares, and the lasso and elastic net selected 14, 13, and 39 base-learner predictions and initial probability estimates from stage one on average, respectively. SCAD rivals the most sparse solution at the meta-level for this dataset and has the highest predictive power.

Table 3.7 displays predictions for the CMEs presented in Tables 3.3 and 3.4. These are extracted from the 10-fold cross-validation during the aforementioned final execution of the *train* function. In stage one, the overall accuracy of the CIGRRF is around 82%[9] for the entire dataset. Within this small sample of data, 14 out of 27 of the CME events are classified correctly. While the CIGRRF does do well in distinguishing the strong CMEs and many of weak ones, it does not identify any of "moderate" CME events or those between -100nT and -200nT here. This demonstrates the difficulty of forecasting using only data near the Sun and the opportunity for improvements to be made. For the second stage, SCAD's predictions are closer to the actual value approximately 59% of the time (16/27) compared to using the best base-learner. More importantly, SCAD succeeds in providing a more optimal prediction for all of the strong CMEs. In fact, it forecasts within -22nT for the dangerous CME emitted on October 28^{th}. Though this is a small sample of data from the entire dataset, it demonstrates the potential advantages

[9]This is obtained from the out-of-bag error estimate from the forest using the *cforestStats* function in **caret**.

Learner	All CMEs		Strong CMEs
(Meta)	WMAE	RMSE	RMSE
SCAD	***18.42***	18.68	***39.04***
Non-negative Least Squares	19.03	***17.51***†	44.33†
Lasso and Elastic Net	19.04	17.57†	45.52†
Linear Regression	19.24	17.70†	45.60†
(Base)	WMAE	RMSE	RMSE
Cubist	20.13†	18.21	47.29†
Extreme Gradient Boosting with Linear Booster	20.49†	19.44	52.06†
Extreme Gradient Boosting with Tree Booster	20.84†	19.31	50.32†
Random Forest	20.99†	18.41	51.24†
Regularized Random Forest	20.99†	18.43	51.38†
Boosted Tree	21.37†	19.15	52.36†
Bayesian Additive Regression Trees	21.57†	20.00†	51.92†
Bagged Multivariate Adaptive Regression Splines	21.85†	19.18	51.98†
Stochastic Gradient Boosting	21.98†	19.50	52.89†
Conditional Inference Random Forest	22.16†	19.01	54.03†
⋮	⋮	⋮	⋮
Quantile Random Forest	26.30†	31.64†	45.35

Table 3.6: Predictive performance for all the meta-learners and the top 10 base-learners (with the quantile random forest). Bold and italics represent the model with the lowest error. The dagger symbol † denotes traditional statistical significance ($\alpha = 0.05$) between SCAD and each learner, respectively.

of stacked generalization for geomagnetic storm prediction.

LASCO Date:Time	*DST*	$\hat{\pi}_{weak}$	$\hat{\pi}_{moderate}$	$\hat{\pi}_{strong}$	SCAD	Cubist
09OCT03:19:33:07	-31	***0.89***	0.11	0.00	***-33***	-22
13OCT03:03:54:05	-77	0.98	0.02	0.00	***-51***	-43
13OCT03:18:30:06	-60	0.89	0.11	0.00	***-53***	-43
15OCT03:00:30:05	-53	0.98	0.02	0.00	***-47***	-45
22OCT03:08:30:32	-49	0.38	0.39	0.23	-71	***-64***
28OCT03:11:30:05	-353	0.23	0.18	***0.59***	***-375***	-432
29OCT03:20:54:05	-383	0.25	0.18	***0.57***	***-290***	-284
03NOV03:19:31:43	-29	***0.53***	0.22	0.26	-58	***-33***
04NOV03:12:06:06	-28	0.23	0.29	0.48	-40	***-33***
04NOV03:12:54:05	-25	0.44	0.45	0.11	-40	***-39***
09NOV03:06:30:05	-62	0.53	0.39	0.09	-82	***-72***
12NOV03:18:30:05	-55	0.99	0.01	0.00	***-46***	-37
13NOV03:22:30:05	-49	***0.91***	0.09	0.00	***-49***	-40
14NOV03:10:54:05	-49	***0.98***	0.02	0.00	***-47***	***-47***
15NOV03:17:50:05	-48	***0.84***	0.15	0.01	-53	***-49***
16NOV03:07:50:05	-42	***0.96***	0.04	0.00	-54	***-49***
18NOV03:08:50:05	-422	0.28	0.22	***0.50***	***-340***	-281
19NOV03:09:26:05	-191	0.50	0.18	0.33	***-147***	-133
19NOV03:15:06:05	-185	0.49	0.24	0.28	***-115***	-103
20NOV03:08:06:05	-105	0.69	0.18	0.13	***-84***	-74
21NOV03:17:50:05	-66	0.80	0.16	0.04	***-62***	-59
21NOV03:19:27:16	-71	0.74	0.22	0.04	***-55***	-49
22NOV03:01:50:05	-47	***0.60***	0.27	0.12	***-63***	-66
23NOV03:09:18:05	-31	***0.90***	0.09	0.01	-52	***-48***
24NOV03:10:50:06	-24	***0.84***	0.13	0.03	-29	***-23***
25NOV03:04:06:05	-22	***0.81***	0.14	0.04	-29	***-24***
30NOV03:12:26:05	-17	***0.80***	0.14	0.06	-35	***-28***

Table 3.7: Hold-out predictions for the sample list of observations in Tables 3.3 and 3.4. For the probability estimates, bold and italics indicate instances where the CIGRRF classified correctly. Note that due to rounding, the probabilities may not add up to one. For the DST predictions, bold and italics denote which model is closer to the observed value: the best meta-learner (SCAD) or the best base-learner (cubist). Ties are denoted by emphasizing both.

3.4.2 Important CME Characteristics

Using only information taken at a CME's initial launch, the variable importance

scores from the CIGRRF can be found in Figure 3.3. Note that since the CIGRRF performs implicit variable selection, not all the predictor variables are listed [47]. The most significant predictor variable for classifying CMEs is the daily average of the solar radio emissions the day the CME is launched [58]. Its high ranking makes sense since this is closely tied with the number of sunspots, which affects the frequency of CMEs during the 11 year solar cycle [36] [75] [47]. In addition, while many full halo CMEs can produce strong geomagnetic storms, less emphasis on AW is indicative that not all full halo CMEs generate these instances [97] [75].

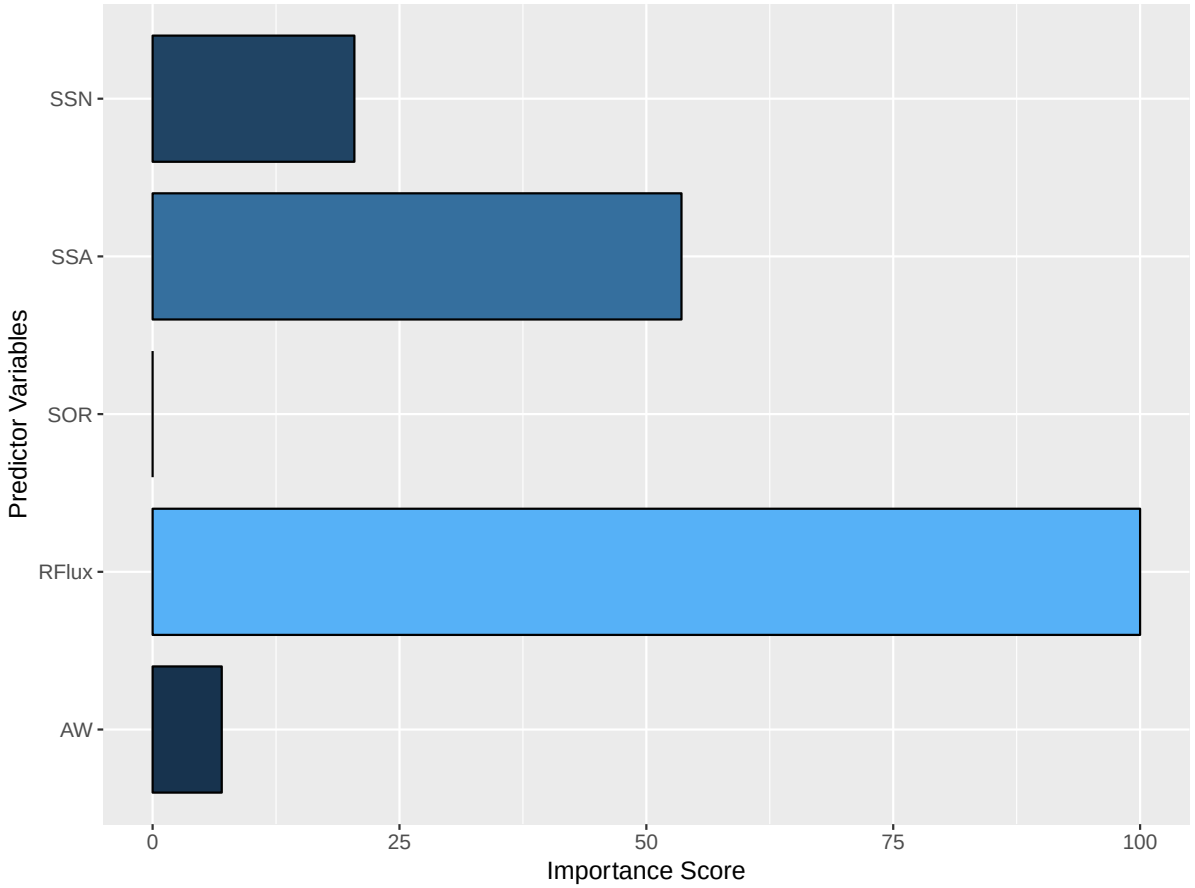


Figure 3.3: Plot of variable importance scores from the CIGRRF using only the stage one predictor variables. The blue bar indicates the importance of each predictor variable in the final estimation on the full dataset. Values range from zero to 100 with 100 signifying the most important.

Following the variable importance procedure described earlier, Figure 3.4 contains a plot of the variable importance scores from stacked generalization. The two most dominating in this list are E_y and B_z. Given the strong relationship between these vari-

ables and the DST value throughout the literature, their contributions towards prediction makes sense. More importantly, the higher value placed on this IPI and lower values on those such as D_p and T_p in determining geomagnetic storm intensity is consistent with other literature (see [74] [24] [94] [38] [40] [47] and references therein). Note that when the IPI information is introduced, the influence of the stage one predictor variables decreases. This is to be expected given the advantages of using IPI.

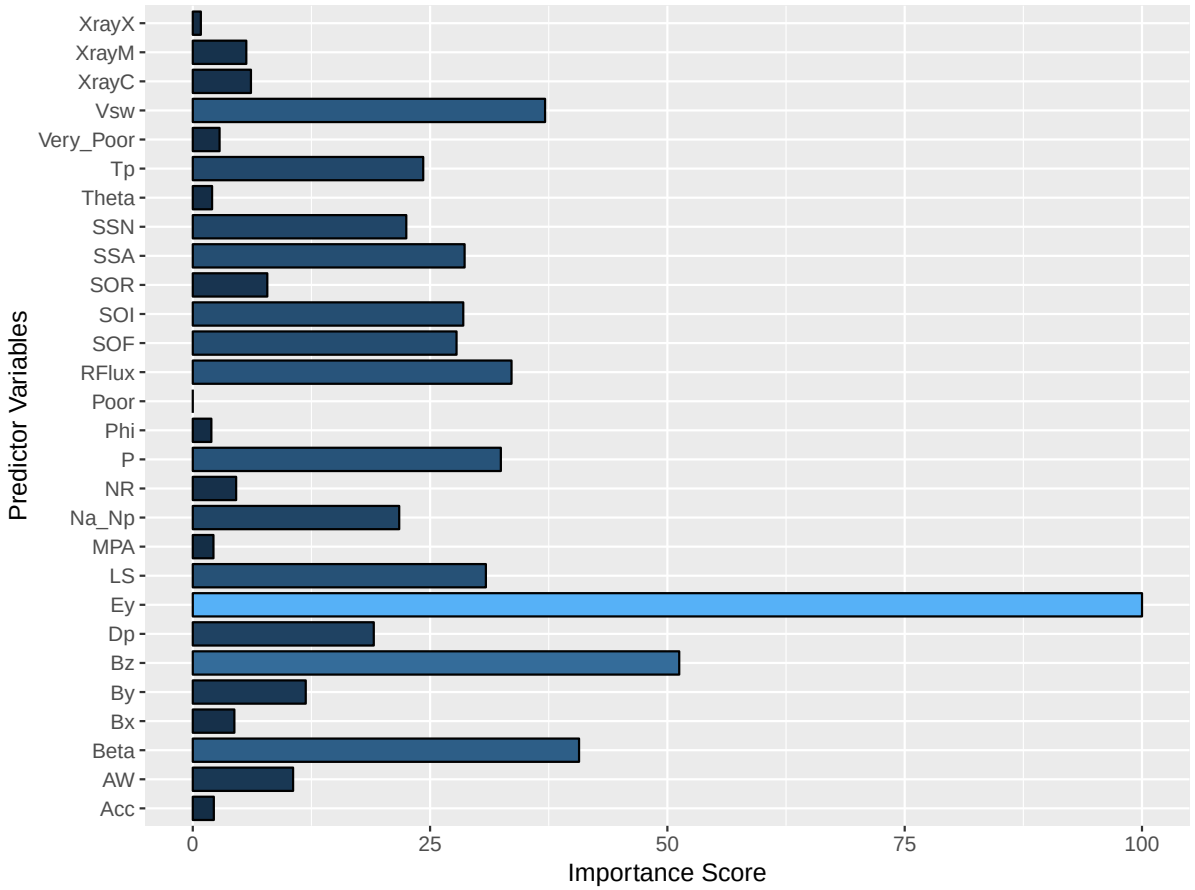


Figure 3.4: Variable importance scores for stacked generalization. The blue bar indicates the importance of each predictor variable in the final estimation on the full dataset. Values range from zero to 100 with 100 signifying the most important.

Table 3.8 reports the six base-learners chosen at the base-level from SCAD to construct the final variable importance scores.[10] Note that all but one are constructed from the top performing base-learners. The only exception is the quantile random forest as it only ranks as the 20^{th} most accurate on this dataset (WMAE = 26.30). On the other

[10]Note that the five predictions from the CIGRRF included at the meta-level (probabilities of generating weak, moderate, and strong geomagnetic storms; entropy; maximal probability) are not selected when SCAD is fit on all of the data; thus, they are not listed here.

Base-learner	$\boldsymbol{\beta}$
Cubist*	0.5131
Quantile Random Forest	0.1670
Extreme Gradient Boosting with Linear Booster*	0.1467
Extreme Gradient Boosting with Tree Booster*	0.1337
Model Tree	0.1173
Random Forest*	0.0002

Table 3.8: The six base-learners used by SCAD to construct the final set of variable importance scores ranked in descending order . Asterisk "*" denotes base-learners with model-specific variable importance measures.

hand, it is the second most vital base-learner. This illustrates the necessary diversity to make meaningful predictions. Each of these base-learners analyzes the prediction problem differently, thereby, allowing SCAD to select the best weights and number of non-zero base-learner predictions to include in the final prediction for a specific region of the conditional distribution of the DST value.

3.5 Discussion

In this work, a meta-learning framework is suggested to predict geomagnetic storms. This approach consists of two stages:

1. Estimate the severity of a CME in question with probabilities using the data available at its launch

2. Update the prediction with a forecasted DST value after collecting the vital IPI via stacked generalization

The general outline is similar to the process by Kim et al. [40]. However, instead of estimating DST in the initial stage, this framework estimates probabilities related to geoeffectivness. Treating stage one as a classification task can help offer a more simplistic interpretation compared to DST, which is likely to dramatically change as the CME propagates through the interplanetary medium. Attaching these probabilities, as well as some information describing the distribution, increases the feature space for the meta-learner. Then, capitalizing on the advantages of stacked generalization, a more intelligent

forecast of DST can be made by utilizing predictions from a wide range of modern techniques. Instead of focusing on the estimation of the conditional mean for DST, quantile regression is implemented at the meta-level to find a better balance between predicting dangerous geomagnetic storms effectively without rendering estimation for the weaker ones useless. By regularizing this meta-learner, it is possible to reduce the size of the metadata and adequately deal with the correlation. Using a regularized quantile regression model at the meta-level provides more adaptability since it can specify specific parts of the conditional distribution and choose the best number of base-learners for that particular region. The posited method is evaluated on an inclusive dataset consisting of various characteristics about the solar wind condition, CMEs, and the Sun. In addition, careful experimental methodology is implemented to estimate generalization error and statistical significance. Results revealed that this framework performs significantly better on the most informative error metrics than the best tuned model or algorithm at the base-level. Moreover, this approach provides an opportunity to study the critical space weather indicators at the beginning of a CME's life and right before its impact on Earth.

Though the study of stacked generalization is not a new concept, this idea has not been explored in the realm of forecasting geomagnetic storm strength from CMEs much if at all. Given the importance of making distinctions, it becomes all the more important to leverage the best analytical tools for space weather prediction. As shown in other studies, it is necessary to incorporate the IPI since these are the most useful for determining the DST. However, as emphasized by Kim et al. [40], this leaves little time to prepare on Earth once the information is collected at the L1 Lagrangian point. Research in attaining the IPI sooner is currently being done. Savani et al. [69] are working towards resolving this type of issue by predicting the magnetic structure of impending CMEs. More accurate forecasts of the IPI will lead to predictions with more lead time. In the meantime, similar to Kim et al. [40], this framework yields a prediction that can be made days out (stage one) and then updates more granularly hours before the final impact (stage two).

Since time is such a factor, computationally efficient approaches must be used. Although stacked generalization requires extra computation, especially for large datasets,

it can be easily parallelized across many clusters since creating the metadata is an independent process. This allows for scalability as new models and algorithms are constantly being developed. Incorporating a larger number of faster and smarter base-learners provides the opportunity to increase predictive power. In addition, the framework needs only to be trained once, setting the parameters in each learner according to historical data. Then, when a new CME is detected, predictions can be made as the new information is gathered with little computational effort.

While work has been done to uncover variable importance schemes in other black box approaches such as in artificial neural networks [60], little work has been done to alleviate some of the "black art" grievances shrouding stacked generalization. By relating the significance of a base-learner's prediction to its own variable importance scores, multiple base-learners can play a part in casting their vote for identifying the relevance of each predictor variable. Offering a way to uncover the influences of the predictor variables at the base-level helps shed light onto the traditional ambiguity surrounding the inner workings of stacked generalization. Naturally, this assumes that in fact predictive performance is tied directly with accurately assessing variable importance, which may not always be the case. In addition, as shown in this work, it can be difficult to find a large number of models or algorithms with model-specific metrics. Restricting the compilation of base-learners to include only those that have model-specific ones may lead to decreases in the much needed diversity of base-learners. Ideally, it would be most beneficial for each base-learner to have its own unique variable importance scores in order to have as many different ways to look at importance as possible, despite whether the base-learner inherently has a model-specific approach. This could be done by using methods that can be applied to any learner, such as the approach taken by Cortez, Cerdeira, Almeida, Matos, and Reis [17].

This study brings several future work opportunities. Firstly, as more data is collected on CMEs in more advanced ways, implementation on larger datasets is possible for both classification and regression tasks. With more data, stacked generalization is more probable to find predictive improvements [81]. In addition, Table 3.7 demonstrated the poten-

tial for stage one predictions to be suboptimal. Looking for ways to improve these are of valuable interest since they give more time to prepare than in the second stage. Secondly, this work only includes 50 base-learners. Increasing this number by incorporating different models and algorithms could yield even better results. Additionally, analyzing the variable importance scores at different quantiles may reveal some new behaviors regarding the predictor variables, much like in quantile process regression [45]. Thirdly, while empirically the proposed framework delivers the best performance, a theoretical treatment as to why this method performs better than any single model is omitted. Future work can be to provide theoretical motivations regarding the use of quantile regression at the meta-level, even if large sample theory does not hold [16], for predicting rare geomagnetic storms. Lastly, introducing some type of cost matrix, as done for MetaFraud [1], or the re-weighting of WMAE can better optimize parameters at both the base and meta-levels.

3.6 Conclusion

Given our dependence on telecommunications and commercial satellites, any disruption in these services could cost millions of dollars for corporations and government agencies worldwide. At the same time, logistically, these entities cannot simply shut down power or communication operations every time a CME approaches Earth. Therefore, it is imperative to make accurate classifications and forecasts as to which of these CMEs that approach Earth can have the potential to trigger devastating geomagnetic storms. Putting into action more sophisticated modeling techniques like stacked generalization has the opportunity to improve, or at least match, the performance when using a single model or algorithm, no matter the dataset. The added benefit of utilizing more complex systems provides the ability to make more accurate predictions, thereby, saving money and reducing the probability for severe geomagnetic storm events wreaking havoc on modern society.

3.7 References

[1] Ahmed Abbasi, Conan Albrecht, Anthony Vance, and James Hansen. Metafraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, 36(4):1293–1327, 2012.

[2] T Andriyas and S Andriyas. Relevance vector machines as a tool for forecasting geomagnetic storms during years 1996–2007. *Journal of Atmospheric and Solar-Terrestrial Physics*, 125:10–20, 2015.

[3] Tohru Araki. A physical model of the geomagnetic sudden commencement. *Solar wind sources of magnetospheric ultra-low-frequency waves*, pages 183–200, 1994.

[4] Lloyd's , Atmospheric and Environmental Research Inc. Solar Storm Risk to the North American Electrical Grid, 2013. Retrieved from `https://www.lloyds.com/~/media/lloyds/reports/emerging%20risk%20reports/solar%20storm%20risk%20to%20the%20north%20american%20electric%20grid.pdf` [accessed: 2016-10-02].

[5] Francesco Audrino and Lorenzo Camponovo. Oracle properties and finite sample inference of the adaptive lasso for time series regression models. *arXiv preprint arXiv:1312.1473*, 2013.

[6] Volker Bothmer and Rainer Schwenn. The interplanetary and solar causes of major geomagnetic storms. *Journal of Geomagnetism and Geoelectricity*, 47(11):1127–1132, 1995.

[7] Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in Knowledge Discovery and Data Mining*, pages 3–12. Springer, 2004.

[8] Pavel Brazdil, Christophe Giraud Carrier, Carlos Soares, and Ricardo Vilalta. *Metalearning: Applications to Data Mining*. Springer Science & Business Media, 2008.

[9] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

[10] Rande K Burton, RL McPherron, and CT Russell. An empirical relationship between interplanetary conditions and DST. *Journal of Geophysical Research*, 80(31):4204–4214, 1975.

[11] Brian S Cade and Barry R Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420, 2003.

[12] HV Cane and IG Richardson. Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, 108(A4), 2003.

[13] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 18. ACM, 2004.

[14] Joseph M Caswell and Nicolas Rouleau. Simple binary prediction of daily storm-level geomagnetic activity with solar winds and potential relevance for cerebral function. *International Letters of Chemistry, Physics and Astronomy*, 17, 2014.

[15] Philip K Chan and Salvatore J Stolfo. Experiments on multistrategy learning by meta-learning. In *Proceedings of the Second International Conference on Information and Knowledge Management*, pages 314–323. ACM, 1993.

[16] Victor Chernozhukov. Extremal quantile regression. *Annals of Statistics*, pages 806–839, 2005.

[17] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[18] Houtao Deng. Guided random forest in the RRF package. *arXiv preprint arXiv:1306.0237*, 2013.

[19] Houtao Deng and George Runger. Feature selection via regularized trees. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

[20] Houtao Deng and George Runger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483–3489, 2013.

[21] M Dryer, Z Smith, CD Fry, W Sun, CS Deehr, and S-I Akasofu. Real-time shock arrival predictions during the "Halloween 2003 epoch". *Space Weather*, 2(9), 2004.

[22] James W Dungey. Interplanetary magnetic field and the auroral zones. *Physical Review Letters*, 6(2):47, 1961.

[23] Saso Džeroski and Bernard Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3):255–273, 2004.

[24] E Echer, WD Gonzalez, and BT Tsurutani. Interplanetary conditions leading to superintense geomagnetic storms (DST <= -250 nT) during solar cycle 23. *Geophysical Research Letters*, 35(6), 2008.

[25] Donald H Fairfield and LJ Cahill. Transition region magnetic field and polar magnetic disturbances. *Journal of Geophysical Research*, 71(1):155–169, 1966.

[26] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[27] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[28] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. *caret: Classification and Regression Training*, 2016. R package version 6.0-76.

[29] Walter D Gonzalez and Bruce T Tsurutani. Criteria of interplanetary parameters causing intense magnetic storms (DST <-100 nT). *Planetary and Space Science*, 35(9):1101–1109, 1987.

[30] N Gopalswamy, S Yashiro, G Michalek, G Stenborg, A Vourlidas, S Freeland, and R Howard. The SOHO/LASCO CME catalog. *Earth, Moon, and Planets*, 104(1-4):295–313, 2009.

[31] Nat Gopalswamy, Alejandro Lara, Seiji Yashiro, Mike L Kaiser, and Russell A Howard. Predicting the 1-AU arrival times of coronal mass ejections. *Journal of Geophysical Research: Space Physics (1978–2012)*, 106(A12):29207–29217, 2001.

[32] JT Gosling, SJ Bame, DJ McComas, and JL Phillips. Coronal mass ejections and large geomagnetic storms. *Geophysical Research Letters*, 17(7):901–904, 1990.

[33] Douglas M Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

[34] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

[35] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[36] Tim Howard. *Coronal Mass Ejections: An Introduction*, volume 376. Springer Science & Business Media, 2011.

[37] Eun-Young Ji, Y-J Moon, N Gopalswamy, and D-H Lee. Comparison of DST forecast models for intense geomagnetic storms. *Journal of Geophysical Research: Space Physics*, 117(A3), 2012.

[38] Eun-Young Ji, Y-J Moon, K-H Kim, and D-H Lee. Statistical comparison of interplanetary conditions causing intense geomagnetic storms (DST <= -100 nT). *Journal of Geophysical Research: Space Physics (1978–2012)*, 115(A10), 2010.

[39] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

[40] R-S Kim, Y-J Moon, N Gopalswamy, Y-D Park, and Y-H Kim. Two-step forecast of geomagnetic storm using coronal mass ejection and solar wind condition. *Space Weather*, 12(4):246–256, 2014.

[41] JH King and NE Papitashvili. Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. *Journal of Geophysical Research: Space Physics (1978–2012)*, 110(A2), 2005.

[42] Roger Koenker. Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89, 2004.

[43] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[44] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[45] Roger Koenker and Jose AF Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999.

[46] Max Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[47] Taylor Larkin. A tree ensemble for classifying geoeffective coronal mass ejections, 2016. Working paper.

[48] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 15. SIAM, 1995.

[49] Michael LeBlanc and Robert Tibshirani. Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91(436):1641–1650, 1996.

[50] Youjuan Li and Ji Zhu. L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 2012.

[51] CA Loewe and GW Prölss. Classification and mean behavior of magnetic storms. *Journal of Geophysical Research: Space Physics*, 102(A7):14209–14213, 1997.

[52] Denise McManus, Houston Carr, and Benjamin Adams. Wireless on the precipice: The 14th century revisited. *Communications of the ACM*, 54(6):138–143, 2011.

[53] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[54] Ronald T Merrill. *Our Magnetic Earth: The Science of Geomagnetism*. University of Chicago Press, 2010.

[55] Ga-Hee Moon. Variation of Magnetic Field (By, Bz) Polarity and Statistical Analysis of Solar Wind Parameters during the Magnetic Storm Period. *Journal of Astronomy and Space Sciences*, 28(2):123–132, 2011.

[56] Dave Mosher and Andy Kiersz. A 100-year solar storm could fry our power grids these are the places most at risk. Retrieved from `http://www.businessinsider.com/solar-storm-risk-map-united-states-2016-9` [accessed: 2016-09-21].

[57] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: A second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[58] National Oceanic and Atmospheric Administration. F10.7 cm radio emissions. Retrieved from `http://www.swpc.noaa.gov/phenomena/f107-cm-radio-emissions` [accessed: 2016-10-05].

[59] National Oceanic and Atmospheric Administration. Index of /pub/warehouse. Retrieved from `ftp://ftp.swpc.noaa.gov/pub/warehouse` [accessed: 2016-09-28].

[60] Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1):135–150, 2002.

[61] Bo Peng and Lan Wang. An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24(3):676–694, 2015.

[62] Christian Petersohn. *Temporal Video Segmentation*. Jörg Vogt Verlag, 2010.

[63] Antti Pulkkinen, Sture Lindahl, Ari Viljanen, and Risto Pirjola. Geomagnetic storm of 29–31 october 2003: Geomagnetically induced currents and their relation to problems in the swedish high-voltage power transmission system. *Space Weather*, 3(8), 2005.

[64] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[65] Sam Reid and Greg Grudic. Regularized linear models in stacked generalization. In *Multiple Classifier Systems*, pages 112–121. Springer, 2009.

[66] IG Richardson and HV Cane. Near-Earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009): Catalog and summary of properties. *Solar Physics*, 264(1):189–237, 2010.

[67] Niall Rooney, David Patterson, and Chris Nugent. Pruning extensions to stacking. *Intelligent Data Analysis*, 10(1):47–66, 2006.

[68] Guzman Santafe, Iñaki Inza, and Jose A Lozano. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508, 2015.

[69] NP Savani, A Vourlidas, A Szabo, ML Mays, IG Richardson, BJ Thompson, A Pulkkinen, R Evans, and T Nieves-Chinchilla. Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 1. Initial architecture. *Space Weather*, 2015.

[70] Rainer Schwenn. Space weather: the solar perspective. *Living Reviews in Solar Physics*, 3(1):1–72, 2006.

[71] Ben Sherwood and Adam Maidman. *rqPen: Penalized Quantile Regression*, 2016. R package version 1.5.1.

[72] Joseph Sill, Gábor Takács, Lester Mackey, and David Lin. Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*, 2009.

[73] Space Studies Board and others. *Severe Space Weather Events: Understanding Societal and Economic Impacts: A Workshop Report*. National Academies Press, 2008.

[74] N Srivastava. A logistic regression model for predicting the occurrence of intense geomagnetic storms. In *Annales Geophysicae*, volume 23, pages 2969–2974, 2005.

[75] Nandita Srivastava and P Venkatakrishnan. Solar and interplanetary sources of major geomagnetic storms during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, 109(A10), 2004.

[76] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):1, 2008.

[77] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

[78] Davor Sudar, Bojan Vršnak, and Mateja Dumbović. Predicting coronal mass ejections transit times to Earth with neural network. *Monthly Notices of the Royal Astronomical Society*, 456(2):1542–1548, 2016.

[79] Masahisa Sugiura. Hourly values of equatorial DST for the IGY. *Ann. Int. Geophys. Yr.*, 35, 1964.

[80] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[81] Kai Ming Ting and Ian H Witten. Issues in stacked generalization. *J. Artif. Intell. Res.(JAIR)*, 10:271–289, 1999.

[82] Ljupčo Todorovski and Sašo Džeroski. Combining classifiers with meta decision trees. *Machine Learning*, 50(3):223–249, 2003.

[83] Chih-Fong Tsai and Yu-Feng Hsu. A meta-learning framework for bankruptcy prediction. *Journal of Forecasting*, 32(2):167–179, 2013.

[84] Bruce T Tsurutani and Walter D Gonzalez. The interplanetary causes of magnetic storms: A review. *Washington DC American Geophysical Union Geophysical Monograph Series*, 98:77–89, 1997.

[85] Jean Uwamahoro, Lee-Anne McKinnell, and John Bosco Habarulema. Estimating the geoeffectiveness of halo CMEs from associated solar and IP parameters using neural networks. *Annales Geophysicae-Atmospheres Hydrospheres and Space Sciences*, 30(6):963, 2012.

[86] Fridrich Valach, Josef Bochníček, Pavel Hejda, and Miloš Revallo. Strong geomagnetic activity forecast by neural networks under dominant southern orientation of the interplanetary magnetic field. *Advances in Space Research*, 53(4):589–598, 2014.

[87] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[88] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006.

[89] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18(2):77–95, 2002.

[90] YM Wang, PZ Ye, S Wang, GP Zhou, and JX Wang. A statistical study on the geoeffectiveness of Earth-directed coronal mass ejections from March 1997 to December 2000. *Journal of Geophysical Research: Space Physics (1978–2012)*, 107(A11):SSH–2, 2002.

[91] Ian H Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[92] David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[93] Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, pages 801–817, 2009.

[94] Yu I Yermolaev, M Yu Yermolaev, IG Lodkina, and NS Nikolaeva. Statistical investigation of heliospheric conditions resulting in magnetic storms: 2. *Cosmic Research*, 45(6):461–470, 2007.

[95] Vasyl Yurchyshyn, Valentyna Abramenko, and Durgesh Tripathi. Rotation of white-light coronal mass ejection structures as inferred from LASCO coronagraph. *The Astrophysical Journal*, 705(1):426, 2009.

[96] Gabriele Zenobi and Padraig Cunningham. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In *European Conference on Machine Learning*, pages 576–587. Springer, 2001.

[97] J Zhang, KP Dere, RA Howard, and V Bothmer. Identification of solar sources of major geomagnetic storms between 1996 and 2000. *The Astrophysical Journal*, 582(1):520, 2003.

[98] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1):239–263, 2002.

[99] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[100] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

4

USING CONVOLUTIONAL NEURAL NETWORKS TO IMPROVE THE INITIAL CLASSIFICATION OF GEOEFFECTIVE CORONAL MASS EJECTIONS

4.1 Introduction

4.1.1 Coronal Mass Ejections and Deep Learning

One of the most significant phenomena to study with regards to space weather is CMEs. These events are incredibly powerful expulsions of magnetic field and plasma from the Sun. They can travel at extraordinary speeds and contain billions of pounds of solar material. When Earthward directed, these explosions can be extremely dangerous and damage many of Earth's technological functions, since they are primarily responsible for causing major geomagnetic disturbances. Hence, it is imperative to study their effects and predict their potential strength before they impact Earth (see [42] [43] and the references therein for a full treatment of the details regarding their effects and strategies for predicting them). Fortunately, satellites are in place to capture data regarding these events that can be exploited to build predictive models and algorithms. However, a prominent issue when predicting CME-driven geomagnetic storms empirically lies in the type of data used. As CMEs travel through the interstellar medium, where they are typically referred to as ICMEs, they interact with the neighboring solar winds and much of their composition can change. This means that utilizing data from the beginning of a CME's life at the Sun is not as useful for predictive purposes as data collected once the CME is much closer to Earth.[1] Yet, a trade-off exists: using the latter type of data leads to very short alert times (usually hours) as opposed to the former which delivers

[1]Larkin [43] noted this consequence in his posited two-stage meta-learning framework for predicting geomagnetic storms.

forecasts days in advance [38]. Therefore, efforts to improve the predictive accuracy of models using data collected at the onset of a CME are vital.

In addition to the main characteristics of CMEs such as their speed and angular width, their images are recorded using the LASCO aboard the SOHO satellite [22]. This instrument produces coronagraph images that obstruct the brilliance of the Sun, similar to a total solar eclipse, in order to more clearly capture activity at the solar corona. The solar corona is the aura of plasma surrounding the Sun, where CMEs are released. An example of one of the types of images that can be produced by the LASCO, courtesy of the NASA and ESA SOHO mission, can be found in Figure 4.1. This CME was the first of several powerful CMEs that caused black-outs in Sweden [51]. Because traditional

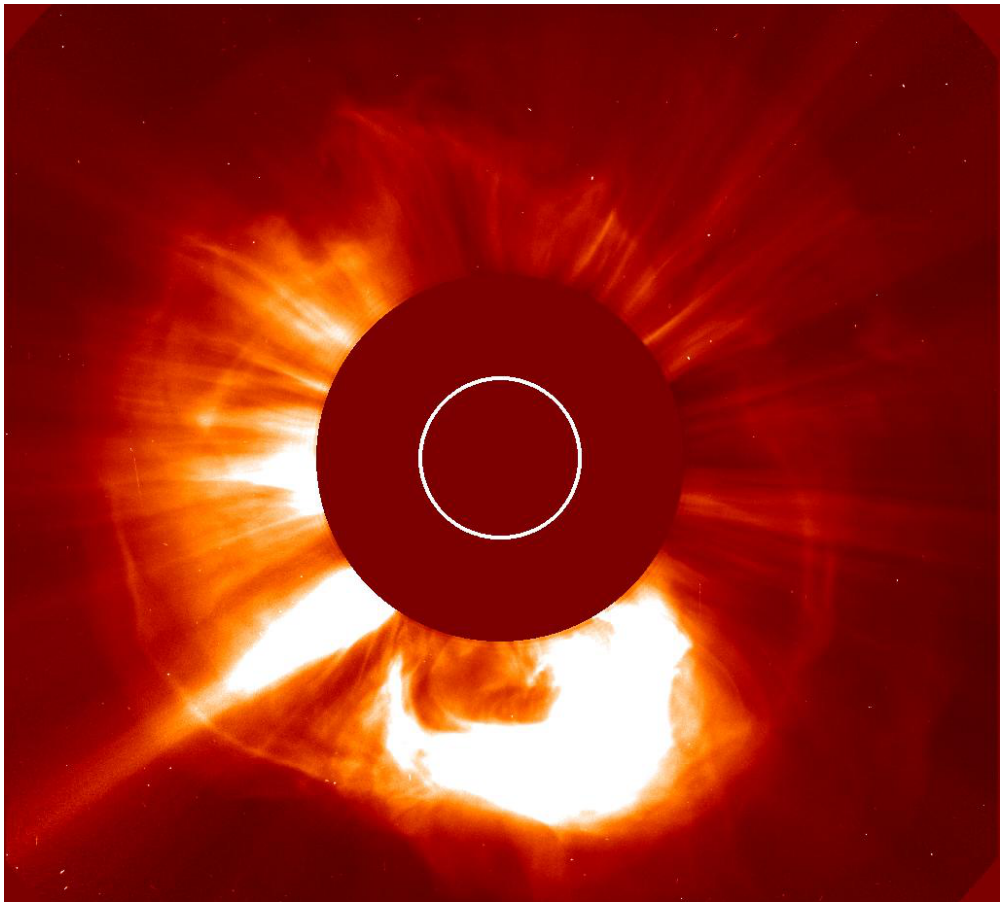


Figure 4.1: Image of a strong CME that occurred October 28^{th}, 2003 at 11:30:05 universal time.

CME detection is done by human observation using these images, a plethora of works in the literature involve developing automated approaches [56] [52] [73]. Oddly enough, little work has been done in making classifications about the geoffectiveness, or impact on

Earth, from these images as they mainly have been focused on identifying the presence of a CME. One of the only studies that acknowledges this opportunity, conducted by Qu, Shih, Jing, and Wang [52], first outlines an automated algorithm to detect a CME. Then, using handcrafted features (or predictor variables)[2] regarding the images, a support vector machine (SVM) [65] is used to classify whether the CME is strong or not, followed by instituting a simple rule based on a CME's speed to distinguish between if the non-strong CME is moderate or weak. While this innovative method yields favorable classification accuracy, the results are based on the training error, which is not reliable and could indicate overfitting [25], for only 100 CMEs. In addition, their approach requires much preprocessing and generation of handcrafted features. Opportunities exist to allow model based feature generation, specifically through the avenues of deep learning.

Deep learning can be defined as a representation-learning method where the goal is to transform some input data into several different, higher-level abstractions in order to learn very complex functions. They usually use a neural network type of architecture with many layers to construct these functions and make predictions, whether it be for regression or classification tasks. LeCun, Bengio, and Hinton [44] noted that the key component of deep learning is that the "layers of features are not designed by human engineers: they are learned from the data using a general-purpose learning procedure" (pg. 436). In this way, one does not have to spend countless hours creating and testing features, the model does it automatically and often with more precision. While the levels of new features are generally unified for prediction using a fully connected layer, they can also be used as generic feature extractors [58]. Implementing deep learning methods has been largely undeserved in space weather studies and is only now beginning to be used for classifying solar events. Recently, Ma, Chen, Xu, and Yan [48] posited a multimodal deep learning model to classify solar radio bursts. Because they instituted a deep learning approach, the model is able to learn about the interactions and correlations amongst the various frequency channels. In the same way, deep learning can be used to improve predictive performance of initial CME classifications once captured by the LASCO by

[2]Predictor variables are generally referred to as features in the machine learning literature. For consistency, the term features will be used instead of predictor variables throughout this work.

allowing the model to create higher-level abstractions of the images.

While deep learning can improve prediction using images directly, these types of models generally require many training images to deliver adequate results [71]; thus, for small datasets, using them as generic feature extractors is a popular alternative. That is, the features generated by a deep learning method are fed into a subsequent classifier. Since only a small fraction of CMEs are directed towards Earth [30], studying a small number of CME events is a likely scenario and common throughout the literature. Hence, high-dimensional situations (i.e., where the number of features is much larger than the number of observations) can arise given that the dimension of deep learning features can be very large. Fan, Han, and Liu [16] discussed that the two goals of general high-dimensional data analysis are developing methods that can "accurately predict the future observations" and "gain insight into the relationship between the features and response for scientific purposes" (pg. 2). Hence, it is advantageous to develop methods that can reduce the number of features while also achieving similar, or better, predictive performance. Moreover, it is imperative to implement techniques that are computationally efficient, given that high dimensionality can be plagued with large sample size. For geomagnetic storm prediction, it is desirable to reduce the number of deep learning features, since it is likely that only certain features will be useful for classifying CMEs. Mapping the relevant deep learning features to the appropriate CME classification can be done using RFs [6], which are flexible supervised learners that are among the most popular machine learning algorithms in use today.

The analysis in this work is two-fold:

1. Posit an efficient and effective high-dimensional feature selection technique based on RFs.

2. Demonstrate the improvement in predictive performance of initial CME classification by including deep learning features alongside the main characteristics of CMEs.

The subsequent sections of this work read as follows. Section 4.2 provides background information beneficial for outlining the methodology. Section 4.3 visualizes the proposed

method and provides detailed explanations of the simulation and real data studies as well as how the experimental CME dataset is constructed. Section 4.4 presents the results of each experiment. Sections 4.5 and 4.6 conclude with a summary and postulate areas for future work.

4.2 Literature Review

4.2.1 Convolutional Neural Networks

One of the most popular deep learning methods for analyzing images are CNNs. CNNs are a type of deep, feed-forward neural network that are especially effective in image classification or recognition tasks. They are comprised of several stages of convolutional and pooling layers [44]. Within each of these convolutional layers, a set of filters is applied to an image. These filters are generally of a much smaller dimension than the input image. They slide, or convolve, over the area of the input image and compute the dot product between the pixel values of the image and the values in the filters to create a 2-dimensional feature map (commonly referred to as an activation map). Each filter will produce a different 2-dimensional feature map based on the focus of the filter (whether it be exposing a certain edge, a blotch of color, or a type of shape) where a different response value exists in each spatial position from the sliding. The values in the feature maps represent neurons that are mapped back to some local region of the input image as calculated from the dot product operation. That is, each neuron is connected to a set of weights of the same dimension as the filter. For example, if the input image is of the dimension $32 \times 32 \times 3$ (a depth of three to indicate the RGB primary color channels) and the filter size, also referred to as the receptive field, is 5×5, then each neuron will be connected to some local $[5 \times 5 \times 3]$ region of the input image and, thus, will have $5 \cdot 5 \cdot 3 = 75$ computed weights plus one bias term. Note that since the same weights are used for the entire sliding process (i.e., the same weights are shared across the entire sliding process, not recomputed for each local region), far fewer parameters are calculated [34]. It is also common to apply some sort of activation function to the subsequent neurons, such as

converting them to rectified linear units (ReLUs),[3] to increase training time [39].

After, it is conventional to apply a pooling layer, which down-samples from the convolutional layer by computing the maximum or average value from a local region of units in a feature map or set of feature maps. The primary focus of this layer is to reduce some of the dimensionality and merge similar features into one [44]. The process of convolutional and pooling layers are typically repeated many times, stacking each layer on top of one another. With each layer, higher and higher-level abstractions of the original image are generated. In order to make predictions from these, a final set of neurons is introduced, noted as the fully connected layer, that connects back to the activations in the previous layer followed by, for classification tasks, the computation of class scores using a softmax function. The training of the weights in the network can be done via backpropagation and gradient descent, as is common in neural networks.

While the concept of CNNs is not new [45], its renaissance, specifically in regards to image classification, was ignited due to their use in the 2012 ImageNet competition [39]. The goal of this competition is to take over a million images from the internet and classify them into one of 1,000 categories. Thanks to new techniques and increased computing power by exploiting graphical processing units (GPUs), the use of CNNs reduced current error rates almost in half [44]. Since 2012, CNNs continued to see success in the ImageNet competition [70] [63].

The most successful CNN architecture as of the 2015 competition is known as ResNet [26], which was developed by a team from Microsoft. This CNN achieved an astounding 3.6% error rate, which surpasses the approximated base-line for human error (5%) [36]. Unlike previous models, ResNet instituted the idea of deep residual learning where the goal is to fit a residual mapping. That is, if some function $\mathcal{F}(x)$ represents the output from a series of convolutional layers with x signifying the initial input into those layers, the function $\mathcal{F}(x) + x$ is fed into a subsequent block of convolutional layers as opposed to $\mathcal{F}(x)$. Here, information about the input x is introduced before the next series of convolutional layers where typically just $\mathcal{F}(x)$ is used. Stacking several blocks on top

[3] $f(x) = max(0, x)$

of one another delivers a very deep network that continues to make marginally modified versions of the original image. The authors explain that using information from the input x in addition to $\mathcal{F}(x)$ makes the optimization process easier.

Although CNNs are extremely powerful tools, they generally require a large number of training observations and computational resources to become effective and feasible. In situations where sample size is limited, building CNNs from scratch often shows poor generalization [71]. However, several strategies exist to combat limited training observations. One popular approach, noted as transfer learning [35], is to use a CNN pre-trained on the ImageNet dataset as a generic feature extractor for other datasets by removing the last layer (that outputs the final prediction) [58], or from previous layers [3], and using the feature maps as features in a different classifier (like a SVM). In this way, a set of new images can be passed through the full network without needing to update the weights (though they can be used to initialize and fine-tune subsequent CNNs [71]). With regard to the limited number of CMEs to study that impact Earth, this strategy can be used to generate higher-level abstractions of CME images to use as features, alongside the main characteristics, to improve classification at the onset of a CME. Naturally, because running the CME images through a CNN like ResNet is an unsupervised task, no knowledge as to which features are going to be most important for CME classification is established. Given that these images are quite different than those from the ImageNet competition, it is likely that only certain features will be useful. In addition, because of the sample size constraints, this problem can easily become high-dimensional in nature. Hence, employing feature selection techniques can greatly aid in creating better and more timely space weather predictions.

4.2.2 High-dimensional Feature Selection with Random Forests

RFs are an incredibly popular machine learning algorithm for data analysis. They involve building multitudes of classification or regression trees on either bootstrapped or subsamples of the training data, randomly selecting from only a sample of features to use for splitting at each node according to some information gain criteria. They have been used in a variety of fields, such as bioinformatics [19], molecular biology [49], handwritten

digit recognition [4], video segmentation [8], geomagnetic storm prediction [42], and others (see [66] for more examples). These supervised learners are very powerful; however, they traditionally have no means of reducing the feature space. Recently, the idea of using GRRF [14] was proposed for gene selection. These models penalize the information gain in a RF in two ways:

(1) Using the feature importance scores from a normal execution of a RF to penalize each feature

(2) Only splitting on the features that add substantially more information gain while being penalized compared to those used in previous splits as the trees are built sequentially

The goal of GRRFs is to select relevant features and then only keep those that are non-redundant. This method was shown to outperform other feature reduction methods, such as lasso [64], and be more computationally efficient compared to the varSelRF [15] procedure, which is also based on RFs. Improvements have been made on this through use of conditional inference frameworks and permutation based feature importance [42]. Unfortunately, the trees in a GRRF can be highly correlated and cannot be used for parallel computing [14] or can be memory intensive if the conditional inference variant is used [42]. To allow for more flexibility in regularized RFs, the concept of using GRFs [12] was introduced. GRFs only perform the first step in GRRFs, which allows them to be parallelized; thus, the features selected here are expected to be relevant, but not necessarily non-redundant [12]. Results showed that GRFs can be more accurate than their GRRF counterpart.

In addition to prediction, RFs have been used to study feature importance. One strategy for classification tasks is using the Gini importance scheme, which involves calculating the mean information gain across all the trees using the Gini index. While this method does well in determining the important features when the data are all continuous, it has been shown to exhibit bias when the data are of different types or varying number of categories [62]. This is because the Gini index favors splitting on features with more split points (e.g., continuous ones) as opposed to ones with less opportunity (e.g.,

binary). This bias will naturally extend to GRFs, since they use the Gini importance to penalize each feature. Considering many gene expression studies deal with continuous data, the GRRF and GRF approaches are sensible for their intended purposes. However, if these methods are to be applied to data of varying types, or even sparse data where some features may have few non-zero entries, the bias issues will need to be addressed. Similar to the work of Larkin [42], this work offers a simple solution to help alleviate some of the bias: deriving the penalties for each feature from a permutation importance scheme designed for high-dimensional data as opposed to using the Gini importance. The proposed strategy, referred to as permutation guided random forests (PGRFs), is easy to implement, computationally faster than GRFs (and sometimes RFs), and helps combat the bias in RFs. To demonstrate the advantages of PGRFs, two simulation and three real data studies are performed.

4.2.3 Guided Random Forests

Define $\boldsymbol{X}$ to be the predictor matrix of dimension $n \times p$ and $\boldsymbol{Y}$ to be the vector of class labels of dimension $n \times 1$ for n observations and p features. Denote (x_i, y_i) as an observation for $i = 1, ..., n$ and x_j as a feature for $j = 1, ..., p$. In particular,

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Furthermore, let *ntree* signify the number of trees used and *mtry* be the number of randomly selected features to choose from at each node v in constructing a RF. To calculate the Gini importance, define the Gini index at each v as

$$Gini(v) = \sum_{c=1}^{C} \mathcal{P}_c^v (1 - \mathcal{P}_c^v) \tag{4.1}$$

where C is the number of classes and $\mathcal{P}_c^v$ is the proportion of observations belonging to class c at node v [42]. To calculate the information gain, the weighted average of the two

descendant nodes v^L and v^R is subtracted from the parent node v. Thus, the more pure the descendant nodes are, the more information gained at the parent node. Therefore,

$$Gain(x_j, v) = Gini(x_j, v) - [\theta_L Gini(x_j, v^L) + \theta_R Gini(x_j, v^R)] \tag{4.2}$$

where θ_L and θ_R represent the proportion of observations appointed to the left and right descendant nodes v^L and v^R, respectively, from the binary split. The x_j with the largest $Gain(x_j, v)$ is utilized for splitting. Define Imp_j as the importance score corresponding to feature x_j. Thus,

$$\text{Imp}_j = \frac{\sum_{v \in S_{x_j}} Gain(x_j, v)}{ntree} \tag{4.3}$$

where S_{x_j} is the collection of nodes in which x_j is calculated to have the largest $Gain(x_j, v)$.

For a sufficiently large $ntree$, it is likely that all the features will be selected at some point in the forest for a split. On the other hand, for a sufficiently large p, a RF may not have the opportunity to split on every feature. In either case, no implicit feature selection takes place. To encourage a sparser solution, Deng [12] proposed regularizing the information gain

$$Gain_R(x_j, v) = \lambda_j Gain(x_j, v) \tag{4.4}$$

where

$$\lambda_j = (1 - \gamma) + \gamma \hat{\text{Imp}}_j \quad \text{such that} \quad \gamma \in [0, 1] \tag{4.5}$$

and

$$\hat{\text{Imp}}_j = \frac{\text{Imp}_j}{max_{j=1}^{p} \text{Imp}_j} \quad \text{such that} \quad 0 \leq \hat{\text{Imp}}_j \leq 1 \tag{4.6}$$

The normalized Gini importance scores $\hat{\text{Imp}}_j$ represent the penalties assessed to each feature's information gain and the importance coefficient γ controls the severity of the penalty. Note that when $\gamma = 0$ this simply reduces down to an ordinary RF as opposed to when $\gamma = 1$ where the maximum regularization is applied. The latter results in using the smallest set of features for splitting.

4.2.4 Permutation Feature Importance

Another approach to assessing feature importance is using the permutation impor-

tance scheme. This involves comparing the prediction errors made by the trees before and after a feature is randomly shuffled, or permuted. The idea is to break the association between the features in $\boldsymbol{X}$ and the response $\boldsymbol{Y}$ by the random shuffle. Hence, if the feature is of great importance, the difference between the prediction error before and after the permutation will be large. Typically, the OOB data (the data not selected from the bootstrapping or subsampling) is used for the permutation in each individual tree. Formally, it can be calculated as so:

$$\text{FI}_j = \frac{1}{ntree} \sum_{t=1}^{ntree} \frac{1}{|\mathcal{B}^{(t)}|} \sum_{i \in \overline{\mathcal{B}}^{(t)}} [I(y_i = \hat{y}_i^{(t)}) - I(y_i = \hat{y}_i^{(t)'})] \tag{4.7}$$

where

- $\overline{\mathcal{B}}^{(t)}$ represents the OOB sample for tree t such that $t \in \{1, ..., ntree\}$
- $\hat{y}_i^{(t)}$ $\{\hat{y}_i^{(t)'}\}$ is the predicted class label for the i^{th} observation before {after} permutation for tree t
- $|\overline{\mathcal{B}}^{(t)}|$ is the number of observations in the OOB sample for tree t
- $\sum_{i \in \overline{\mathcal{B}}^{(t)}} [I(y_i = \hat{y}_i^{(t)}) - I(y_i = \hat{y}_i^{(t)'})]$ is the sum of the difference in accuracy before and after permutation for tree t where $I(\cdot)$ is an indicator function

Features that are unimportant will have a value near zero or even negative, due to randomness. This approach has been shown to be a much safer option to assess feature importance because of the bias issues related to the Gini index [62].

However, the permutation importance (as well as the Gini importance) does not establish any sort of threshold denoting significance. That is, no universal rule exists to decipher how positive a feature's importance score needs to be in order to be considered truly influential. Research has been done to establish an appropriate cut-off to separate the signal from the noise. For example, Altmann, Toloşi, Sander, and Lengauer [1] offered a heuristic that outputs a p-value[4] from the permutation importance scores. It involves constructing a null distribution from several RF iterations built on versions of the data

[4]It should be noted that p-values regarding feature importance are not strict mathematical p-values since it is unclear whether the referenced population parameter even exists [32].

where the response is permuted and calculating the respective importance scores. A p-value can be created, either parametrically or non-parametrically, by comparing these back to an original set from a RF where the response is not permuted. However, this can be very computationally demanding for high-dimensional datasets because it requires multiple RF executions (anywhere between 50 and 100 is recommended). In light of this, Janitza, Celik, and Boulesteix [32] posit a faster approach where the null distribution is approximated using the non-positive importance scores calculated from two-fold cross-validation and reflecting these about zero to establish symmetry. That is, instead of using the OOB observations for the permutation, a hold-out sample is used such that

$$\mathrm{FI}_j^{(h)} = \frac{1}{2}\sum_{k=1}^{2}\frac{1}{ntree}\sum_{t=1}^{ntree}\frac{1}{|S_k|}\sum_{i\in S_k}[I(y_i = \hat{y}_i^{(t)}) - I(y_i = \hat{y}_i^{(t)'})] \tag{4.8}$$

where

- S_k represents the sample of data *not* used in tree construction (the hold-out sample)
- $|S_k|$ is the number of observations in the hold-out sample

In this way, two RFs are built: the first on S_2 and the other on S_1. Then S_1 is used to calculate the importance scores from the first RF while S_2 is used for the second. This results in permutation importance scores that are completely independent of one another, which are then averaged together. The full feature importance test can be summarized as the following:

Step 1: Randomly partition the training data into two disjoint sets. Construct a RF on each disjoint set and calculate Eq. 4.8

Step 2: Approximate the null distribution from the union of the following three sets

1. $A_1 = \{\mathrm{FI}_j^{(h)} | \mathrm{FI}_j^{(h)} < 0 \quad \forall \quad j = 1, ..., p\}$
2. $A_2 = \{\mathrm{FI}_j^{(h)} | \mathrm{FI}_j^{(h)} = 0 \quad \forall \quad j = 1, ..., p\}$
3. $A_3 = -A_1$

Step 3: Extract the p-value for each feature using the the empirical cumulative distribution function $\hat{F}_0$ of the null distribution. That is, compute $1 - \hat{F}_0(\text{FI}_j^{(h)})$.

The product is a set of importance scores and corresponding p-values for each feature that reflects a monotonic relationship: the higher the importance score, the lower the p-value. The authors showed that using the OOB permutation importance (Eq. 4.7) to approximate the null results in a skewed distribution which may not preserve Type I error; hence, this hold-out version should be preferred. While their approach is computationally much faster than that of Altmann et al. [1], a caveat exists: it requires a large number of non-positive importance scores to appropriately construct a sufficient null distribution. Thus, it performs best in high-dimensional situations where many of the features will be uninformative. Fortunately, within the context of this work, it is likely that many non-positive importance scores will be observed since only certain feature maps derived from the CNN will be useful for classifying CMEs.

4.2.5 Permutation Guided Random Forests

To institute PGRFs, one can simply substitute the use of Gini importance in Eq. 4.6 with that from the permutation importance scheme described in the previous section for high-dimensional data. Because the approach is able to quantify p-values, a preliminary feature selection step can take place based on the rejection level α. The main goals of this step are to increase computational speed by significantly reducing the feature space prior to executing the GRF and to eliminate much of the initial noise. Then, a second round of feature selection can occur using the penalized information gain on the remaining features to ensure only the most influential are being chosen. Formally, the new information gain can be defined as

$$Gain_R(x_j, v) = \lambda_j Gain(x_j, v) \quad \text{such that} \quad x_j \in \hat{\text{FI}}_j^{(h)} \tag{4.9}$$

where

$$\lambda_j = (1 - \gamma) + \gamma \hat{\text{FI}}_j^{(h)} \quad \text{such that} \quad \gamma \in [0, 1] \tag{4.10}$$

and

$$\hat{\text{FI}}_j^{(h)} = \frac{\{\text{FI}_j^{(h)} | 1 - \hat{F}_0(\text{FI}_j^{(h)}) < \alpha \quad \forall \quad j = 1, ..., p\}}{max_{j=1}^{p}\{\text{FI}_j^{(h)} | 1 - \hat{F}_0(\text{FI}_j^{(h)}) < \alpha \quad \forall \quad j = 1, ..., p\}} \tag{4.11}$$

such that $0 \leq \hat{\text{FI}}_j^{(h)} \leq 1$ and $\alpha \in (0, 1]$. Note that α has an inverse relationship with γ: as α approaches zero, more features will be screened out and vice versa as α approaches one. Naturally, only the features that are not screened out in the first round will have a penalty, hence, $x_j \in \hat{\text{FI}}_j^{(h)}$ in Eq. 4.9. Because of the potential high-dimensionality of the feature maps, instituting two different feature selection techniques in the same model can greatly aid performance. Because these are both RF based, the proposed method benefits from the flexibility that tree models offer, especially considering the possibility for interactions and other non-linear effects when the number of features is large. The improvement of PGRF over GRF as a general feature selection method will be shown via simulation and real data studies in the subsequent sections. Combining this approach with deep learning to enhance initial geoeffective CME classification will be shown as well using real CME data.

4.3 Methodology

All experiments are conducted in the R environment for statistical computing [53] version 3.3.3 on a personal computer with an Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz processor and 16 GBs of RAM. RFs and GRFs are implemented using the *RRF* function from the **RRF** package [13] [14] [12]. The permutation importance strategy for high-dimensional data proposed by Janitza et al. [32] is implemented using the *holdoutRF* function from the **ranger** package [69]. Deng [12] found that building RFs using only the features selected by GRF are more accurate than using the GRF as classifiers themselves. Thus, a final execution of a RF on the selected features from each method occurs in all applicable experiments except the second simulation study, since this is not evaluating error. To take advantage of the ability to parallelize on the real and CME data studies, each call of the *RRF* function is run in parallel across the eight core processor using the **foreach** package [2]. Note that this involves only one call for the RF, three calls for

the GRF (importance scores, feature selection, final prediction), and two calls for PGRF (second round of feature selection and final prediction).

To remain consistent with Deng [12], the same parameter settings are applied for both simulation and real data studies ($ntree = 1000$, $\gamma = 1$, everything else at default). The rejection level $\alpha = 0.1$ in the PGRF is chosen because this denotes the common threshold for identifying marginal significance. For the CME data study, the final execution of the RF is conducted using the *cforest* function from the **party** package [28] [62] [61]. This RF implementation, noted as a CIRF, provides unbiased feature selection and permutation importance. Additionally, subsampling (i.e., sampling without replacement) is performed instead of bootstrapping (i.e., sampling with replacement) when executing the *RRF* function since the latter can induce biases [62] [32]. Note that this does not apply to the *holdoutRF* function because it uses the aforementioned two-fold cross-validation approach to train the forests. All other arguments in each function are left at their default settings (aside from increasing *ntree* to 1,000).

To evaluate the performance of each model on the real and the CME data studies, five classification metrics are reported: the AUC, PRAUC, LogLoss, accuracy, and the kappa. The first three are based on the predicted class probabilities while the latter two are based on the predicted class labels. AUC has been shown to be a good one number summary of performance [31]. The PRAUC [54] is similar to AUC, except it plots the proportion of predicted positives that are truly positive (precision) against the true positive rate (recall) as opposed to plotting the false positive rate against the true positive rate like in AUC. LogLoss [21] is a calibration score that seeks to calculate how close a probability prediction is to the true class designation [57]. Accuracy is the conventional measure for the correctness in the class labels while the kappa [68] measures the accuracy of a model when juxtaposed against predictions made at random. Each probability based metric is calculated using the one-versus-all approach [40] since some of the experiments have more than two response classes.[5] For all metrics except LogLoss,

[5]Note that in cases where the number of classes is only two, the one-versus-all approach will yield the same results for AUC and LogLoss compared to the typical binary uses of these measures since it does not matter which class is considered the event. However, PRAUC will be slightly different since it does matter. Thus, for binary cases, PRAUC is simply the average from declaring each class as the event.

the larger the value the better. All metrics are estimated using 10-fold cross-validation or 10×10 repeated cross-validation when in the presence of small sample size [37]. The latter is only applied in experiments where $n < 1,000$. Estimating these metrics using both strategies is done by implementing each candidate method as a custom model within the **caret** framework [18]. Furthermore, significance tests between the PGRF and the other candidate models are conducted for each performance criteria (10 estimates in the 10-fold cross-validation and 100 in the repeated counterpart) using the corrected repeated k-fold cross-validation test [5] [57]. This test has been shown to produce acceptable Type I error, low Type II error, and good replicability for comparing two models.

4.3.1 First Simulation Study Set-up

Adopting the code example for GRF shown by Deng [12], a 500×500 data matrix of values from a random uniform distribution is generated. The response is created from the summation of first and twenty-first columns, meaning that these are the only useful features. It is then converted into a binary feature by using the median as a cut-off. Half of the data is used for training while the other is used for testing. The number of features selected and the number misclassified observations on the test set is recorded for a RF, GRF, and PGRF. In addition to these models, it is noteworthy to investigate the performance of the feature selection when the GRF component is removed from the PGRF (i.e., training a RF on features selected from Eq. 4.11 only). This model will be noted as a permutation random forest (PRF). Instituting the PRF as well can illuminate if the second feature selection step is necessary. The experiment is repeated 1,000 times generating a new dataset in each iteration. The purpose of this study is to demonstrate the prediction and feature selection capability of PGRFs in comparison to GRFs, PRFs, and RFs.

4.3.2 Second Simulation Study Set-up

Using a similar experiment performed by Strobl, Boulesteix, Zeileis, and Hothorn [62], 200 features following a normal distribution with a mean of zero and standard deviation of 0.5 are randomly generated alongside 50 following a binomial distribution with the probability of success at 0.5. A standard deviation of 0.5 is used for a more fair comparison

with the simulated binary features [20]. A binary response is created by summing the 50 binary features together and using the median again. Thus, a simulated dataset with 250 features where only the binary ones have any relevance to the response is created. Sample size is set to $n = 120$ [62] to promote a high-dimensional situation. This experiment is repeated 1,000 times with new datasets being generated at each iteration. It should be noted that these simulation settings do not yield an easy classification problem, since the response is constructed from a summation of 50 random binary features. However, since the purpose of this study is to demonstrate that PGRFs have less biased feature selection than GRFs, comparing their relative performance is sufficient.

4.3.3 Real Data Studies Set-up

Three real datasets are selected for analysis: two from the NIPS 2003 Feature Selection Challenge [23] (Arcene and Gisette), and the Amazon Commerce Reviews data [47]. All are classification problems and can be found in the UCI Machine Learning Repository [46]. A summary of the basic information about the datasets can be found in Table 4.1.

Dataset	Real Features	Random Probes	Observations	% Non-zero	Classes
Arcene	7,000	3,000	200	54%	2
Gisette	2,500	2,500	1,000	13%	2
Amazon	10,000	0	1,500	15%	50

Table 4.1: Summary of real datasets.

Arcene is a mass-spectrometry dataset whose task is to distinguish cancer from normal patterns. Both training and validation datasets are combined to created the final dataset (200×10000). Approximately 44% of the observations indicate cancer, and the features are all continuous. Gisette is a handwritten digit recognition problem. Each observation represents the pixels associated with an image of a nine or a four. In order to execute this as a high-dimensional data problem, only the validation set is used (1000×5000). Both classes are equally represented, although the features are quite sparse (only 13% of the entries are non-zero). For these two datasets, the total number of features are composed from real features and random probes. Random probes are features that are drawn from a similar distribution as the real features, but are completely unrelated to the

classes. The Amazon dataset is real-world, writeprint dataset whose goal is identifying online authorships using 10,000 writeprint features encapsulating linguistic behavior such as the use of punctuation, frequency of certain words, length of sentences, etc. Features are based on 30 reviews from 50 authors. Notably, this classification problem is difficult due to the number of classes, high-dimensionality (1500×10000), and sparse nature. No preprocessing steps are applied to Arcene and Gisette. Features in the Amazon dataset are scaled to have a variance of one [67].

4.3.4 CME Data Study Set-up

To select which CME events to study, the catalog compiled by Richardson and Cane [7] [55] is used (noted as the ICME list). This list provides a recording of ICME events and probable CME association from the LASCO catalog [22] related to the respective ICMEs. In addition, their list records the minimum DST value during an ICME's time of effect on Earth. These two qualities are vital to making a connection back to the initial CMEs recorded by the LASCO catalog and establishing the geoeffectiveness of the ICME, respectively. The list of C2 images corresponding to each CME event is queried from the SOHO database at 1024 resolution [59]. The C2 white light coronagraph is used since it captures images closer to the Sun. To obtain an informative sample of images, only those that match the probable LASCO CME date and time recorded in the ICME list exactly are used. Each of these are given a storm classification based on in its recorded DST value [42] [43]. Events with a DST of -100 nT or less are labeled strongly geoeffective. Those with a DST of -50nT or more are labeled weakly geoeffective. Otherwise, it is considered moderately geoeffective.

The main characteristics of a CME are obtained from the LASCO catalog [22]. For simplicity, only those events where the main LASCO catalog information, such as its speed and angular width, is not missing is included for analysis. This results in a list of 160 CMEs spanning from January 1997 through January of 2016 where 45, 58, and 57 are considered weak, moderate, and strong, respectively. The mass and kinetic energy columns are omitted due to the increased presence of missing values and because these are largely uncertain [22]. For any ICME events that share the exact same departure

time, the event with the larger angular width is chosen, since this likely leads to a more significant CME event [72] [60]. Alongside the LASCO catalog, daily solar information from NOAA is also included for the day of ejection [50]. Table 4.2 summarizes the set of CME and Sun characteristics used as features. Table 4.3 displays a sample of the most recent CME events dating back to 2010 included in the analysis.

x	Type	Description
AW	C	Sky-plane width of CME (degrees)
LS	C	Linear speed of CME (km/s)
SOI	C	Quadratic speed of CME at initial height measurement (km/s)
SOF	C	Quadratic speed of CME at final height measurement (km/s)
SOR	C	Quadratic speed of CME at height of 20 solar radii (km/s)
Acc	C	Acceleration of CME in (m/s^2)
MPA	C	Measurement position angle of CME at the height-time measurements (degrees)
$Poor$	B	Noted as a poor event in the comments
$RFlux$	C	Daily average 10.7cm flux values of solar radio emissions on CME ejection day in 10^{-22}J/s/m^2/Hz
SSN	D	Number of sunspots recorded on CME ejection day
SSA	C	Sum of the corrected area of all observed sunspots on CME ejection day in millionths of the solar hemisphere
NR	D	Number of new sunspot regions on CME ejection day
$XrayC$	D	Number of C-class solar flares on CME ejection day
$XrayM$	D	Number of M-class solar flares on CME ejection day
$XrayX$	D	Number of X-class solar flares on CME ejection day

Table 4.2: List of main CME and Sun characteristics to use as features. Feature types are indicated as C-Continuous, D-Discrete, B-Binary.

Similar to other works regarding transfer learning [3] [41], this work explores where in a

LASCO CME MM/DD/YYYY UT	DST	Classification	*AW*	*LS*	*SOI*	*SOF*	*SOR*	*Acc*	*MPA*	*Poor*	*Rflux*	*SSN*	*SSA*	*NR*	*XrayC*	*XrayM*	*XrayX*
02/07/2010 03:54	-7	weak	360	421	415	427	429	0.5	113	0	90	51	360	1	5	1	0
04/03/2010 10:33	-81	moderate	360	668	677	658	661	-1	171	0	77	27	250	0	0	0	0
04/08/2010 04:54	-67	moderate	160	264	293	235	179	-2.2	235	0	76	23	50	0	0	0	0
05/24/2010 14:06	-80	moderate	360	427	367	492	474	3.8	280	0	73	17	90	0	0	0	0
08/02/2011 06:36	-15	weak	268	712	852	554	596	-15.5	285	0	122	98	1255	0	5	1	0
08/04/2011 04:12	-115	strong	360	1315	1539	1074	1208	-41.1	298	0	116	81	1380	1	13	1	0
09/06/2011 23:05	-75	moderate	360	575	561	589	582	1.1	300	0	112	93	560	1	0	1	1
09/14/2011 00:00	-72	moderate	242	408	362	452	457	3.2	302	0	143	144	890	2	8	0	0
09/24/2011 12:48	-118	strong	360	1915	1575	2254	2089	79.6	78	0	190	88	1930	2	4	8	1
01/18/2012 12:24	-8	weak	203	267	274	259	253	-0.5	173	0	148	122	1130	1	3	1	0
03/07/2012 00:24	-131	strong	360	2684	2982	2379	2594	-88.2	57	0	136	102	1800	0	1	0	2
03/13/2012 17:36	-74	moderate	360	1884	1706	2054	1987	45.6	286	0	141	80	650	0	2	1	0
05/12/2012 00:00	-41	weak	360	805	855	755	760	-6.6	107	0	130	85	1190	0	8	0	0
07/02/2012 08:36	-8	weak	360	1074	1265	878	990	-26.9	85	0	166	165	1130	0	17	4	0
07/04/2012 17:24	-68	moderate	360	662	760	553	0	-37.6	124	1	163	129	990	2	11	7	0
07/12/2012 16:48	-127	strong	360	885	680	1092	2265	195.6	158	0	165	132	1750	2	4	0	1
11/09/2012 15:12	-108	strong	276	559	519	601	603	4	157	0	115	65	410	0	1	0	0
03/15/2013 07:12	-132	strong	360	1063	882	1247	1161	25.8	112	0	123	105	650	0	2	1	0
04/11/2013 07:24	-6	weak	360	861	929	792	819	-8.1	85	0	137	121	960	0	7	1	0
07/09/2013 15:12	-73	moderate	360	449	517	382	290	-7.7	174	1	120	98	720	0	4	0	0
10/06/2013 14:43	-62	moderate	360	567	414	710	822	21.5	10	1	107	53	200	0	2	0	0
12/12/2013 03:36	-22	weak	276	1002	1121	876	943	-15.3	236	0	165	156	1140	0	6	0	0
04/18/2014 13:25	-25	weak	360	1203	1130	1279	1245	13.5	238	0	172	263	1840	0	4	1	0
09/10/2014 18:00	-75	moderate	360	1267	1556	950	1119	-51.6	175	0	160	161	1070	0	1	0	1
06/22/2015 18:36	-86	moderate	360	1209	1369	1065	1147	-25.1	358	0	135	77	1320	1	4	1	0
09/18/2015 05:00	-75	moderate	131	823	468	1196	1030	35.5	188	0	103	62	490	0	4	0	0
11/04/2015 14:48	-96	moderate	360	578	502	650	701	10.1	288	0	114	93	580	1	7	3	0
12/28/2015 12:12	-117	strong	360	1212	1182	1243	1228	4.6	163	0	112	64	530	0	4	1	0
01/14/2016 23:24	-104	strong	360	191	162	227	286	2.3	234	0	103	36	260	1	0	0	0

Table 4.3: Sample of the most recent CMEs used for analysis.

CNN yields the best feature maps for prediction. Using a strategy similar to Athiwaratkun and Kang [3], the accuracy of a classifier (a CIRF with features selected by PGRF in this work) is assessed on various layers from a CNN (ResNet-18 pre-trained on the ImageNet dataset [26]). Specifically, the CME images are run through ResNet-18, creating multiple sets of feature maps of differing dimensions depending on the layer. Then, the accuracy of the classifier at each layer is measured using the CME classifications as the response. Note that these are extracted after the initial convolutional layer (conv1), after each subsequent residual block (conv2_x, conv3_x, conv4_x, conv5_x), and before the application of the softmax function [26] (noted as the CNN codes [3]). A summary of the layers used is listed in Table 4.4.

Table 4.4: Layers used from ResNet-18 to extract the feature maps.

Layer	Name	Dimension	Features
Raw Pixels	"data"	$32 \times 32 \times 3$	3,072
Conv_1	"conv0_output"	$16 \times 16 \times 64$	16,384
Conv_2	"_plus1_output"	$8 \times 8 \times 64$	4,096
Conv_3	"_plus3_output"	$4 \times 4 \times 128$	2,048
Conv_4	"_plus5_output"	$2 \times 2 \times 256$	1,024
Conv_5	"_plus7_output"	$1 \times 1 \times 512$	512
CNN Codes	"fc1_output"	1000	1000

The feature maps are extracted using the pre-trained *resnet-18* CNN from the R package **mxnet** [9]. The CME images are resized from 1024×1024 to 32×32 for model training. After extraction, they are $L2$ normalized, as is a common preprocessing step for the feature maps [58].

4.3.5 Integrating Deep Learning Features

To truly evaluate if an improvement exists in using the deep learning features for preliminary CME classification, it is necessary to establish base-line performance. This can be done by executing a classifier on the non-image data (i.e., the main CME and Sun characteristics). Then, after executing the same classifier on the non-image data *and* the feature maps selected from the PGRF, a comparison can be made back to the base-line. Ideally, if the feature maps are predictive, then performance will increase for

that classifier. Three classifiers are tested on the CME data: a CIRF and two SVMs. The SVMs are included because of their popularity in transfer learning [58] [3] and CME classification [52] [10]. The two SVMs are implemented based on the parameter settings from Choi, Moon, Vien, and Park [10] and Qu et al. [52], respectively. Both of these have found success in CME classification. The former uses a radial basis function kernel with a cost penalty of 2^{-5} and decision boundary penalty of 2^{-15} to predict whether or not a CME will be geoeffective. The latter uses a linear kernel to make an initial prediction about whether a CME will be strong or not. Then, with a simple rule regarding their speed, the non-strong CMEs are segmented out: if the LS speed is greater than 300 km/s, then it is considered of moderate strength and weak otherwise [29] [11] (as cited in [52]). Both SVMs are implemented using the **kernlab** package [33] within the **caret** framework. The architecture for adding the deep learning features into initial CME prediction is outlined in Figure 4.2.

In the diagram, the red rectangles represent data, the blue circles symbolize models, and the yellow cloud indicates an operation. Note that the structure is very simple, devoid of complicated preprocessing steps like in the work of Qu et al. [52]. This is thanks to the use of deep learning to create new features instead of relying on handcrafted ones. The PGRF step will significantly reduce the number of feature maps to include in the training dataset, considering that some layers have over 16,000 features.

4.4 Results

4.4.1 First Simulation Study Results

Table 4.5 displays the average number of features selected ($\overline{F}$) and testing observations misclassified (out of 250) on average during the 1,000 iterations. Note that since this experiment is repeated, these values are slightly different for the RF and GRF compared to those reported by Deng [12].

Unsurprisingly, the RF selects every feature in each iteration. Since all but two are irrelevant, it makes sense that utilizing all the available features yields the highest error. The GRF, PRF, and PGRF select a fewer number of features and, thus, achieve

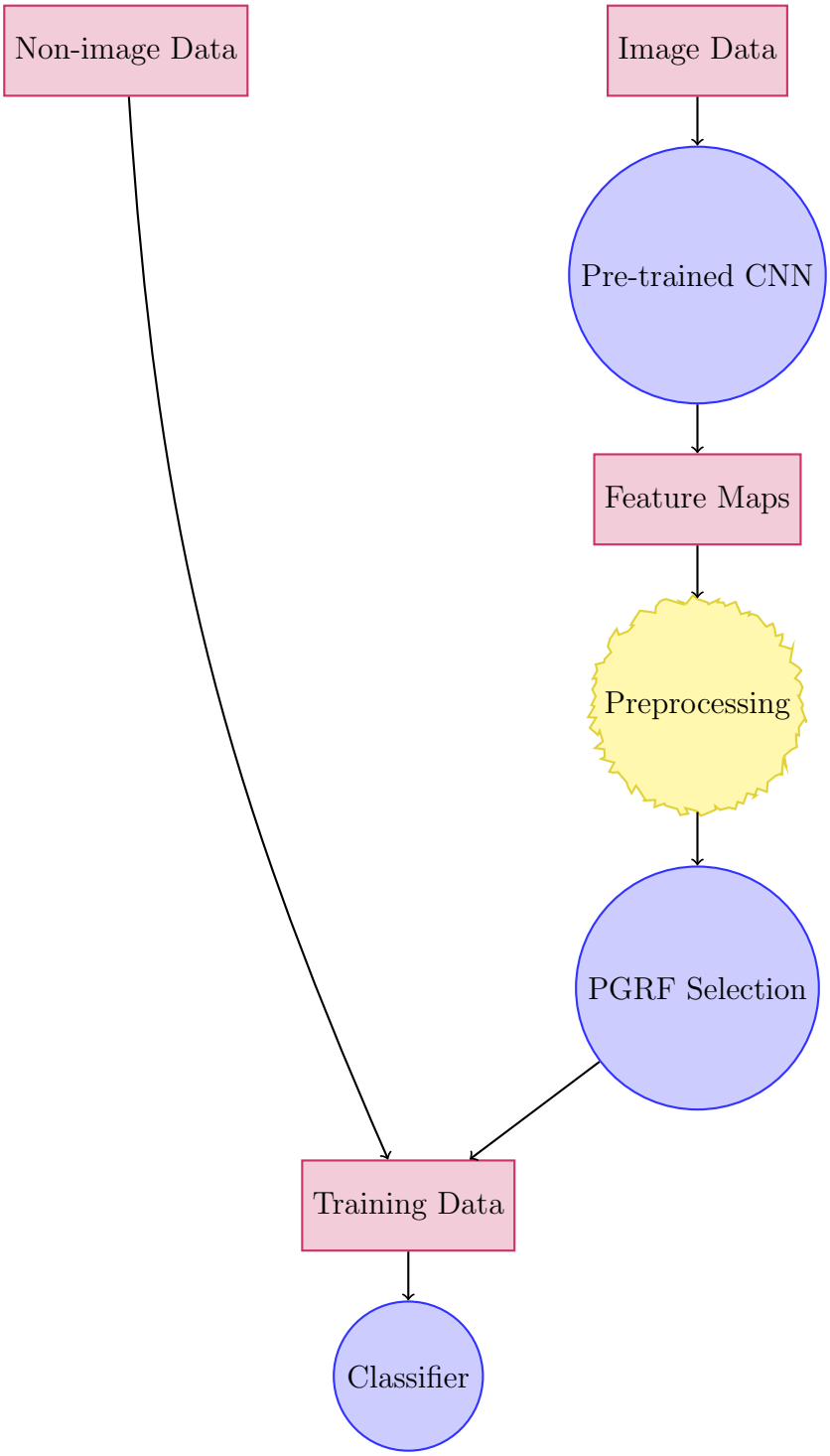


Figure 4.2: Architecture for the proposed approach for improving initial CME classification.

Model	$\overline{F}$	Number Misclassified
RF	500.00	45.60
GRF-RF	212.49	38.63
PRF-RF	51.57	29.37
PGRF-RF	45.72	28.93

Table 4.5: Results on first simulation study.

a smaller number of misclassified observations. While the differences between the PRF and the PGRF may seem minor, note that this is not a true high-dimensional setting. It is likely that in such a scenario, the differences between the number of features selected, as well as the error rate, will become greater. Furthermore, if the true solution is found by the PRF, then the PGRF will yield the same or a very similar conclusion because the most relevant features should still be selected, even while penalized. Given this, and the fact that the proposed PGRF here selects the fewest features and is the most accurate, the PRF will not be considered in any of the subsequent experiments.

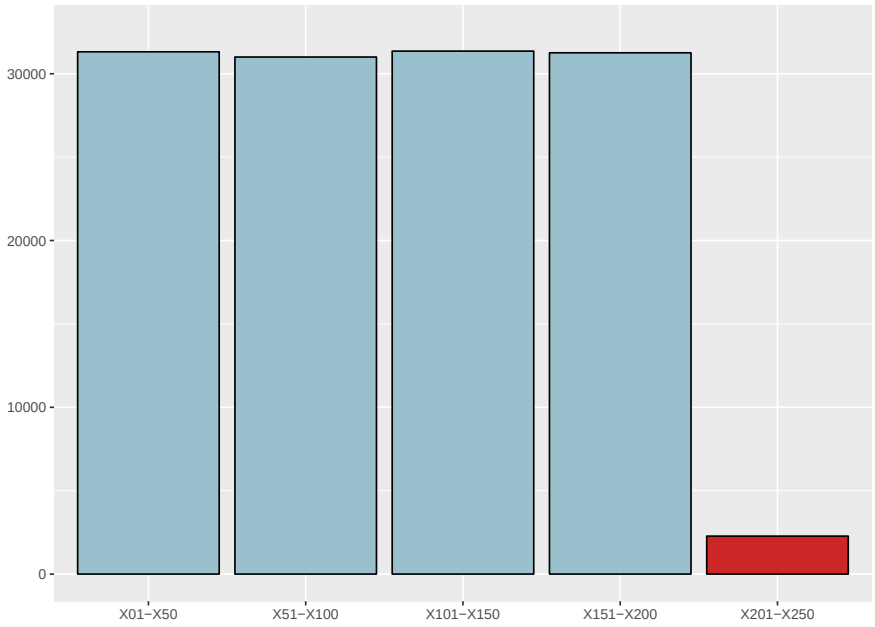
4.4.2 Second Simulation Study Results

Figure 4.3 shows the feature selection frequencies over the course of the 1,000 iterations. For readability, the 250 simulated features are bucketed into groups of 50. The last group of 50, denoted in red, represents the relevant binary features.

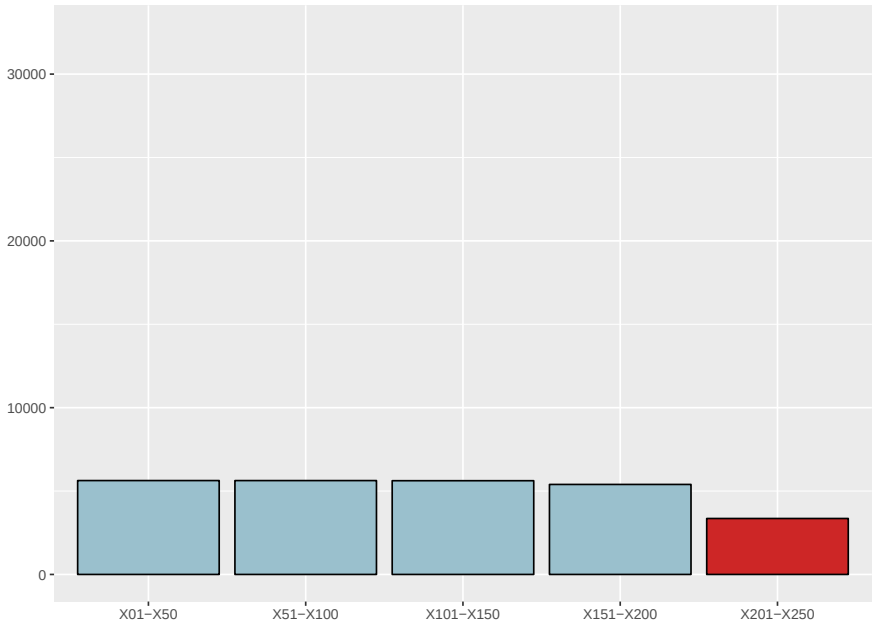
The bias towards the continuous features in the GRF is glaring as it selects these at a much higher frequency compared to the relevant binary ones, regardless of the fact that they are truly uninformative. On the other hand, it is clear that the PGRF selects a greater proportion of the relevant features and far less uninformative ones. Naturally, the PGRF selects a smaller subset of features on average compared to the GRF. This supports the need to address the bias issues in GRFs when studying data of varying types.

4.4.3 Real Data Study Results

Table 4.6 displays the average classification metrics for each real dataset. Bold and italics values represent the best metrics. Dagger symbols "†" indicate when statistical differences between the PGRF and the RF/GRF are found at the conventional 0.05 sig-



(a) $\overline{F} = 127.21$



(b) $\overline{F} = 25.63$

Figure 4.3: Feature selection frequency for (a) GRF and (b) PGRF.

nificance level. In addition to the classification metrics, $\overline{F}$, the average time (in minutes) of model execution within each fold, and summary ranking is also shown. The summary ranking indicates where the PGRF ranked in comparison to the RF and GRF out of a given metric across datasets.

	AUC	PRAUC	LogLoss	Accuracy	Kappa	$\overline{F}$	Time
Arcene							
RF	0.9248	0.8242	0.4195†	0.8364	0.6664	6184.27	0.71
GRF-RF	***0.9406***	***0.8411***	***0.3609***	***0.8546***	***0.7041***	446.45	1.63
PGRF-RF	0.9307	0.8304	0.3727	0.8435	0.6814	***270.07***	***0.57***
Gisette							
RF	0.9877	0.9660	0.2191†	0.9500	0.9000	4201.50	3.02
GRF-RF	0.9893	***0.9604***	0.1773†	***0.9540***	***0.9080***	740.20	7.02
PGR-RF	***0.9899***	0.9560	***0.1704***	0.9530	0.9060	***536.10***	***2.56***
Amazon							
RF	***0.9880***	***0.5369***	0.0577†	0.7913	0.7871	9959.90	***17.14***
GRF-RF	0.9866	0.5237	0.0512†	0.7767	0.7721	931.00	33.63
PGRF-RF	***0.9880***	0.5348	***0.0489***	***0.7927***	***0.7884***	***792.10***	17.23
Rank	1.33	2.33	1.33	1.67	1.67	1	1.33

Table 4.6: Classification results on real datasets.

According to the classification metrics, no method dominates over another in terms of highest/lowest value overall. However, it is important to note that the PGRF significantly outperforms the RF via LogLoss in each dataset and the GRF via LogLoss on Gisette and Amazon. In no instance is the PGRF outperformed with significant difference. Furthermore, is has the fastest time of execution in two of the three datasets while selecting the fewest features amongst all of the candidate methods.

While the initial screening allows the PGRF to be much faster than the GRF, it may seem counter-intuitive for the PGRF to obtain speed faster than the RF due to its multi-step process. However, thanks to the computational fast implementation of RFs on high-dimensional data, the *holdoutRF* function in the **ranger** package runs very fast compared to the *RRF* function. Because the initial screening out of features can occur quickly, the subsequent two executions of the *RRF* function (second round of feature selection and prediction) are much faster since the number of potential features to use at each node in each tree is drastically reduced. A more fair assessment of speed between

the RF and the PGRF could be accomplished by replacing the executions *RRF* function with the RF implementation from **ranger**, where possible. However, in order to remain consistent with the work of Deng [12], this is not done. The most important take-away from these real data studies is the considerable increase in speed of the PGRF compared to the GRF (which is more than twice the time of the RF), which is also backed by selecting the smallest set of features.

The proposed approach appears to perform the best on the difficult Amazon dataset, achieving the best values in each category except PRAUC. Given that this particular dataset is quite sparse, it is likely that the bias towards selecting features with more split points (i.e., features with more non-zero entries) consequently leads to a larger F for the GRF, as seen in the second simulation study. In addition, given that other works have revealed that only a proportion of features are needed for good predictions [47] [67], it is expected that a good approximation of the null distribution can take place for the initial feature selection, which will lead to better regularization in the information gain and, ultimately, better predictions. While the proposed method performs favorably in all three datasets, perhaps it is most effective when studying sparse data with a large number of irrelevant features (as also observed by the increased performance of PGRF over GRF on Gisette compared to Arcene).

While the RF executed on features selected by the PGRF outperforms the other two methods, it does not yield higher 10-fold cross-validation accuracy compared to the synergetic neural network approach proposed by Liu, Liu, Sun, and Liu [47] and the sparse group lasso model posited by Vincent and Hansen [67]. However, this result may be more of a reflection on using a RF to make the final predictions as opposed to the actual feature selection process. To investigate the performance of a different classifier, ridge regression [27] is also tested. This model has advantages over ordinary least squares regression on correlated data, which is a likely consequence in high-dimensional situations. In a multiclass setting, ridge regression solutions can be obtained solving the following

log-likelihood function [24]:

$$\underset{(\beta_{0k},\beta_k)\in\mathbb{R}^{k(p+1)}}{\text{minimize}} -\left\{\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{k=1}^{K} y_{il}(\beta_{0k}+x_i'\beta_k)-\log(\sum_{k=1}^{K} e^{\beta_{0k}+x_i'\beta_k})\Big]+\frac{\lambda}{2}\sum_{k=1}^{K}\sum_{j=1}^{p}\beta_{kj}^2\right\} \quad (4.12)$$

such that β is a $p\times K$ matrix of coefficients for K classes, β_k signifies the k^{th} column, β_j is the j^{th} row, and y_{il} denotes an indicator function as elements in a $n\times K$ response matrix where l represents the class of interest. Because ridge regression does not shrink features all the way to zero, it does not perform feature selection. Hence, it is a natural candidate model to pair with the PGRF. To institute ridge regression for this study, the *cv.glmnet* function from the **glmnet** package [17] is used. Other than specifying the proper family (multinomial) and penalty argument ($alpha = 0$), the function is executed at its default settings. In addition to same parameter settings as from previous experiments, the PGRF is also executed with 2,000 and 3,000 trees. Table 4.7 displays the 10-fold cross-validation accuracy and number of features selected during each fold on average.

	[47]	[67]	PGRF-Ridge	PGRF-Ridge 2000 Trees	PGRF-Ridge 3000 Trees
Accuracy	0.8049	∼0.82	0.8187	***0.8287***	0.8213
$\overline{F}$	2000	∼1000	***805***	934	1016

Table 4.7: Comparison of performance on Amazon dataset. The tilde symbol ∼ denotes approximated values as reported by authors.

Notably, as the number of trees increases in the PGRF, the number of features selected also increases. The monotonic relationship here is plausible since with more trees, more features could potentially be used for splitting. However, more features does not necessarily translate to better accuracy. Performance plateaus when the number of trees is set to 2,000. Combining the PGRF with ridge regression outperforms using the PGRF with a RF as well as the other comparison methods on this dataset. In addition, the PGRF based approach also yields the sparsest solution. These points demonstrate the ability to use PGRF feature selection in conjunction with other classifiers to best suit the prediction problem.

Overall, based on simulation and real data studies, the proposed PGRF provides

competitive predictive performance to that of the RF and GRF while selecting sparser solutions compared to the RF and being more computationally efficient compared to the GRF. The ability to use memory-efficient algorithms and parallelize the processes in the PGRF is particularly important for scalability reasons (e.g., applying this feature selection method to even larger datasets with higher degrees of dimensionality). These results are encouraging for implementing this approach to select the most important deep learning features.

4.4.4 CME Data Study Results

To better understand feature maps from each layer, Figures 4.4 through 4.9 plot the first twenty feature maps as rescaled grayscale images[6] for the strong CME depicted in Figure 4.1. Note that this image is resized to 224×224 [26] before entering ResNet-18 for visualization purposes here.

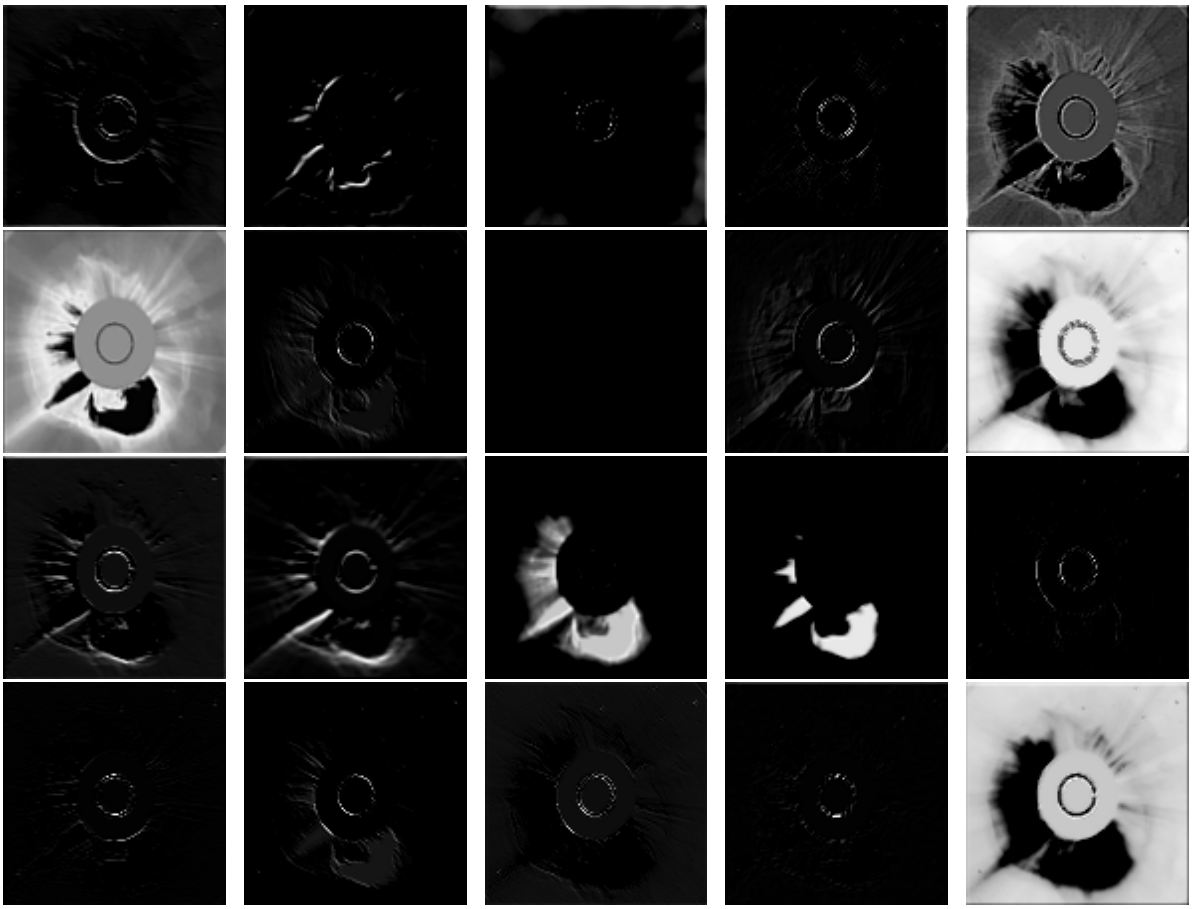


Figure 4.4: First 20 feature maps from conv_1.

[6]In addition, the ReLU activation function is applied so that no negative values exist for the rescaling.

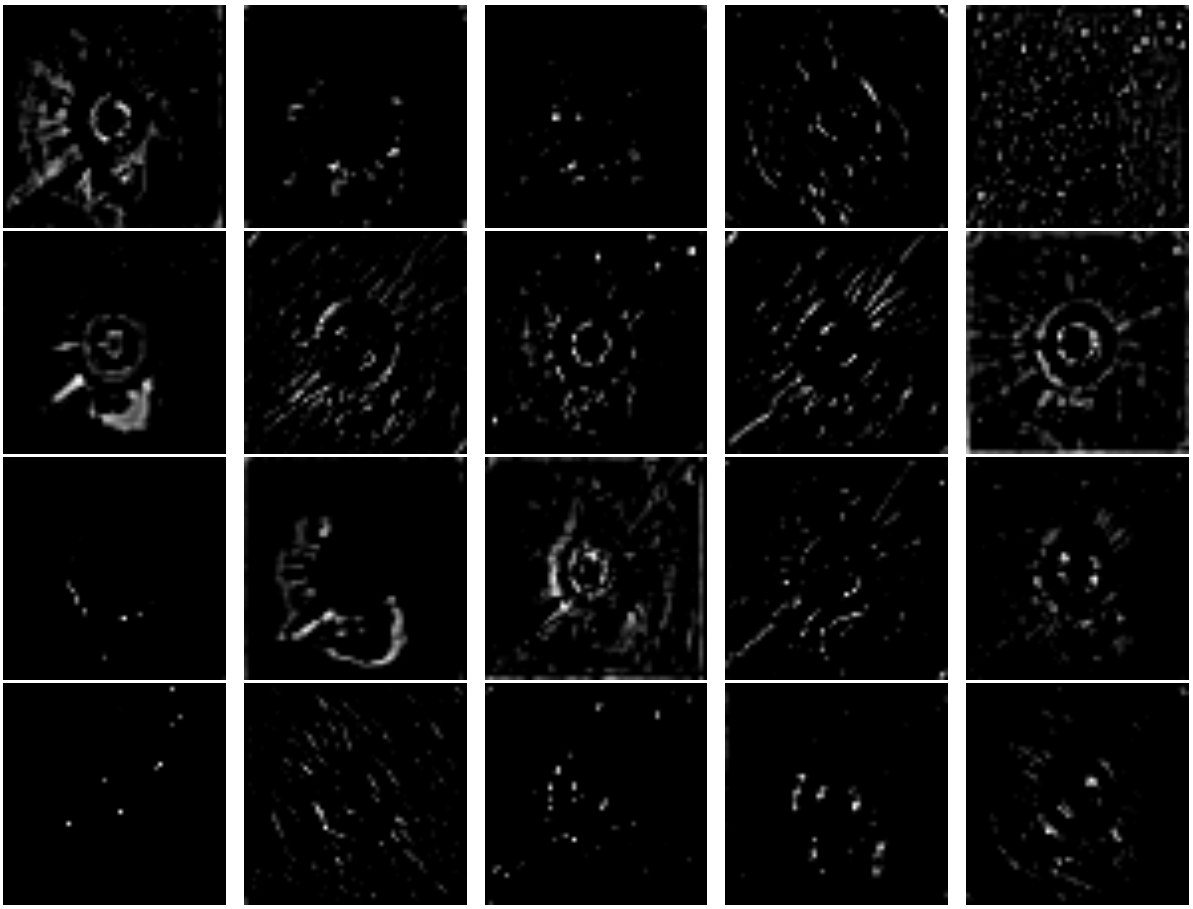


Figure 4.5: First 20 feature maps from conv_2.

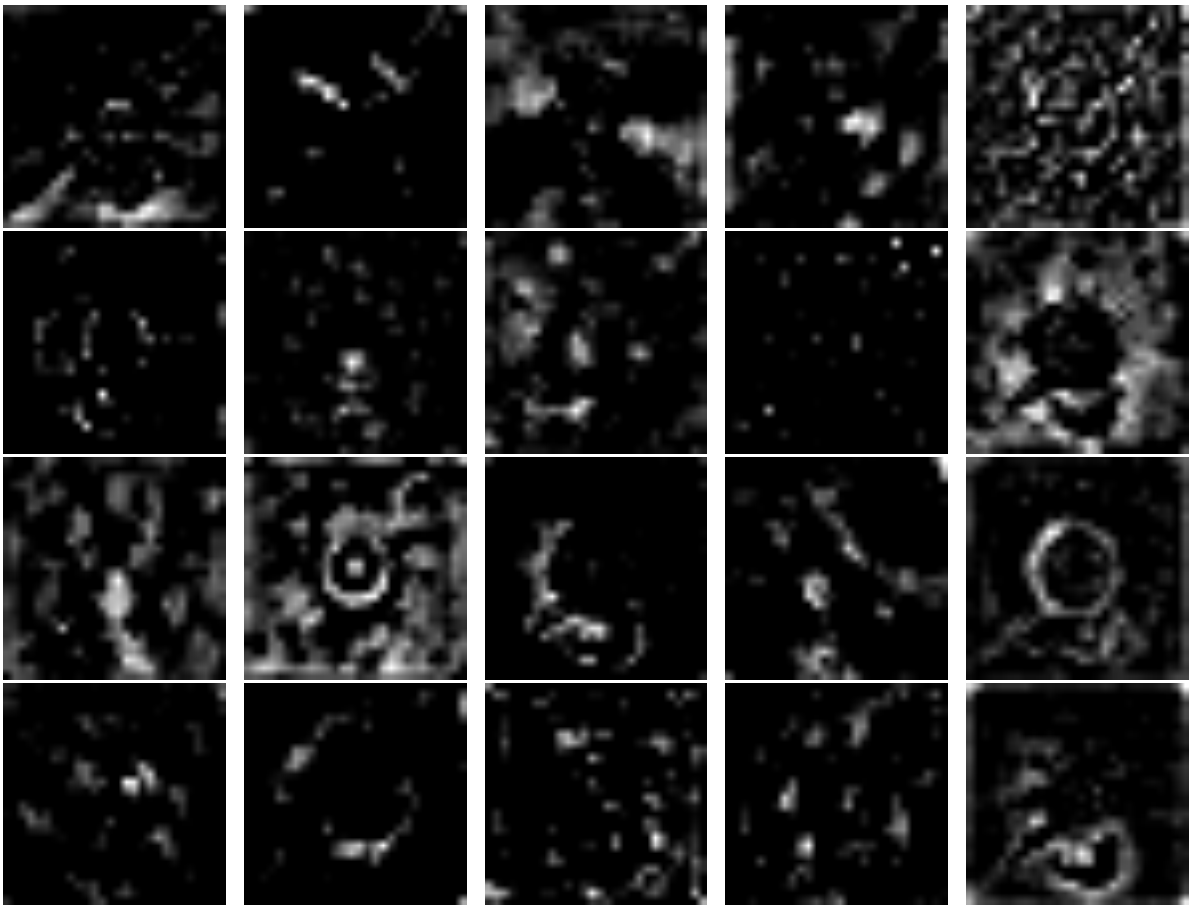


Figure 4.6: First 20 feature maps from conv_3.

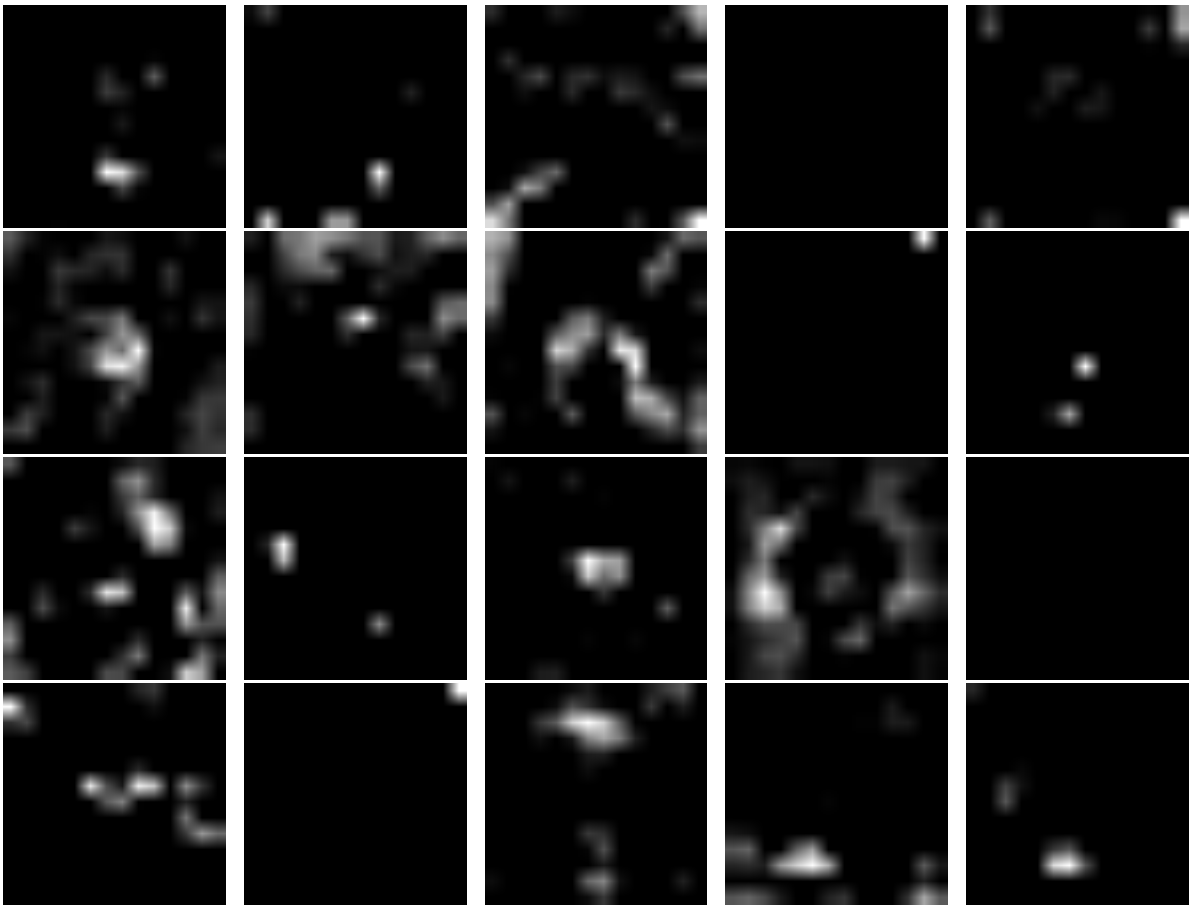


Figure 4.7: First 20 feature maps from conv_4.

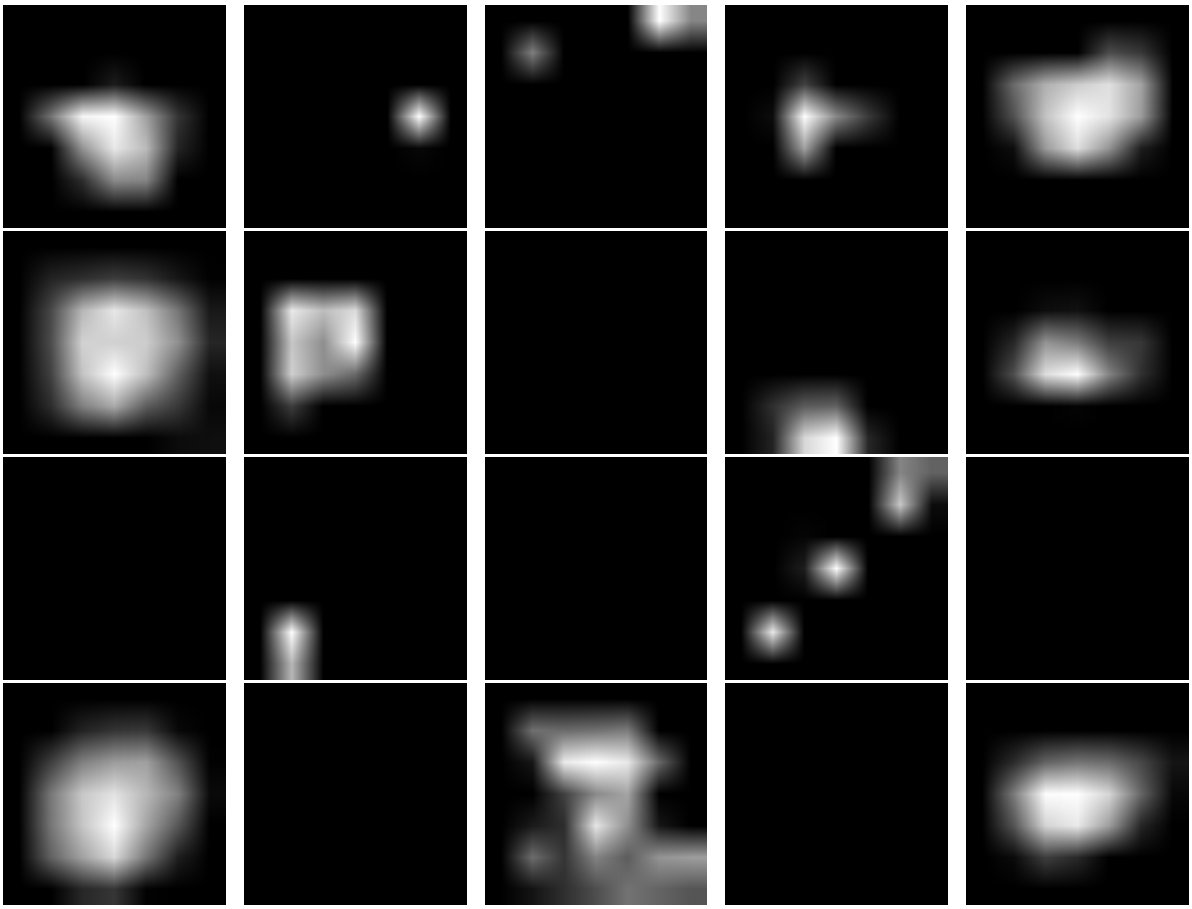


Figure 4.8: First 20 feature maps from conv_5.

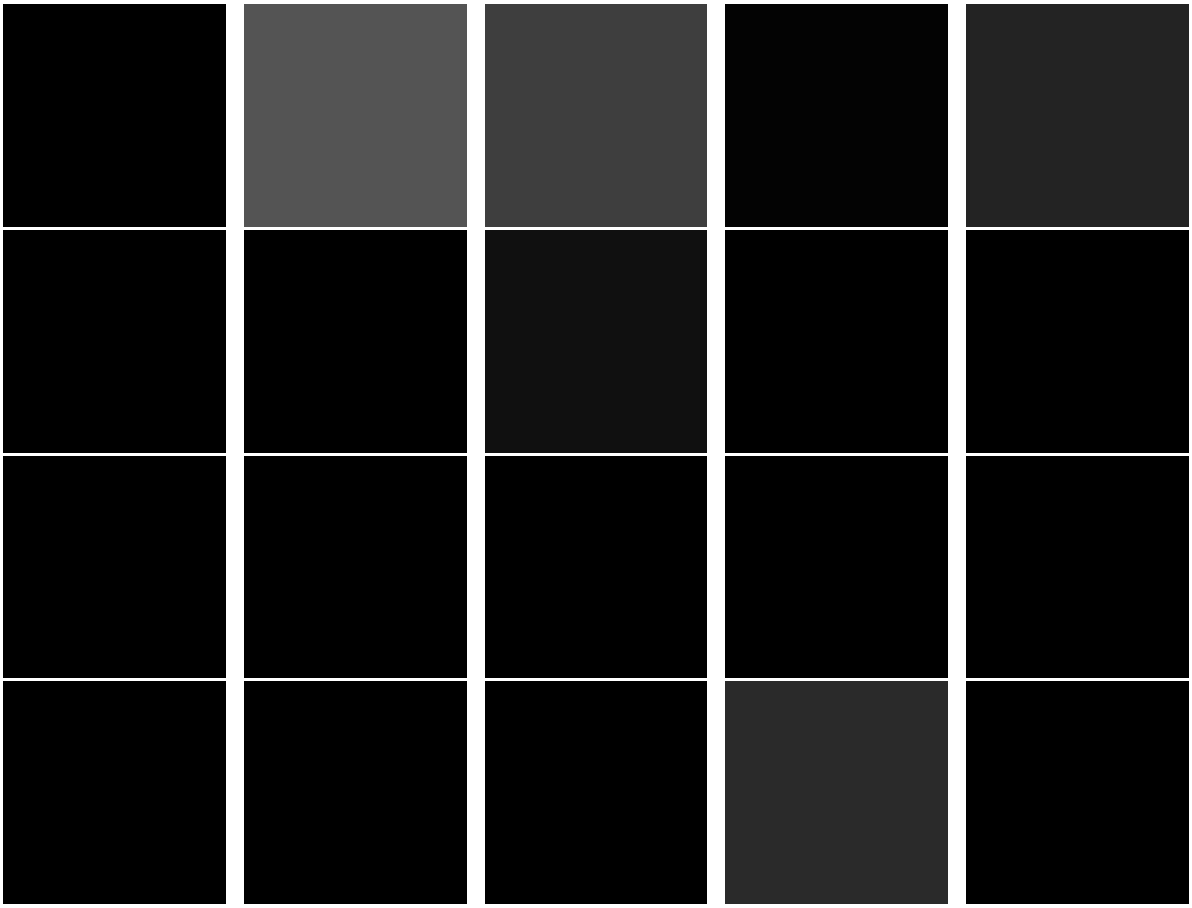


Figure 4.9: First 20 feature maps from the CNN codes. Note that each square represents one value from one feature.

Notably, in the first convolutional layer, the image looks recognizable. Each filter produces a feature map that focuses on a different aspect, whether it be the surrounding solar corona or the actual ejection. However, with each layer, residual mapping distorts the original image until it becomes indistinguishable. In the last layer, the CNN codes indicate a value that belongs to a neuron that activates because this higher-level abstraction is a useful feature for prediction. As expected, many of the squares are black, meaning that the value is negative, and thus, not useful for plotting. This reinforces the hypothesis that only certain feature maps will be useful for making predictions.

Figure 4.10 plots the accuracy of a CIRF with PGRF feature selection when executed on each layer, including the raw pixel values, with the CME classification as the response. As noted by Athiwaratkun and Kang [3], it is possible to find better predictive performance from feature maps extracted earlier in the CNN. Here, the first convolutional layer yields the highest accuracy. The increased error rate in the deeper layers makes sense since the weights become increasingly specific towards the ImageNet competition data. This phenomenon is a likely consequence when the number of desired images is small and very different than those on which the CNN is trained [35]. Hence, only the feature maps from the first convolutional layer will be introduced into the proposed architecture in Figure 4.2.

Table 4.8 shows the classification performance of the classifiers with and without the deep learning features. SVM version are denoted as SVM_{Choi} and SVM_{Qu} corresponding to the parameters set by Choi et al. [10] and Qu et al. [52], respectively. For AUC, PRAUC, and kappa (with statistical difference), the best performing model is the SVM_{Choi} where the additional information about the image is included. For the other two metrics, LogLoss and Accuracy, the CIRF with the integrated deep learning features deliver the best results. In the majority of the metrics, the inclusion of the deep learning features delivers better predictive performance. The only instances where this is not true is when using the Qu et al. [52] approach, which performs poorly here. This is due to the use of the simple classification rule to segment moderate CMEs from weak ones. This causes the probability predictions to be absolutely certain (=1) when the SVM classi-

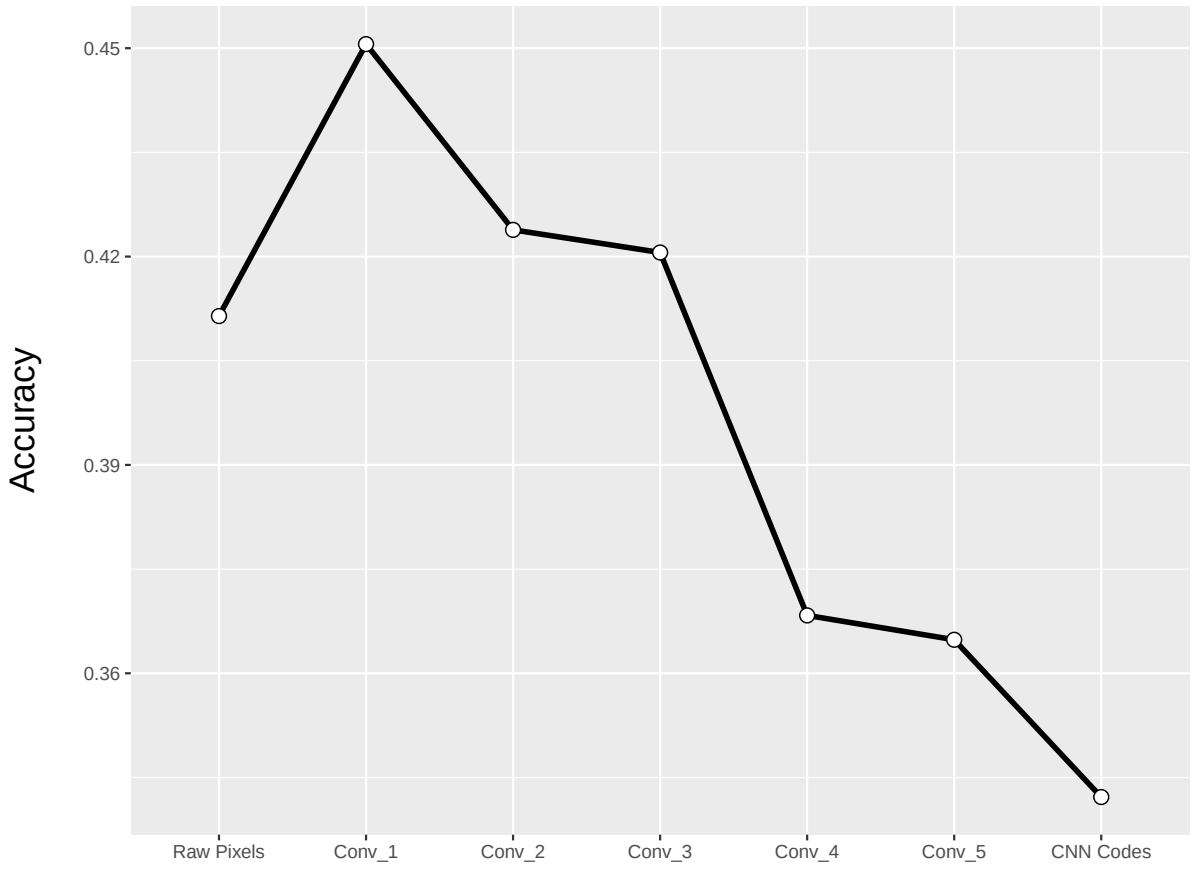


Figure 4.10: Accuracy of PGRF-CIRF across each layer in ResNet-18.

fied an event as not strong. Regardless, the accuracy and kappa are also relatively low, indicating that their method is not as useful on this dataset as it was on theirs.

	AUC	PRAUC	LogLoss	Accuracy	Kappa
CIRF	0.6114	0.3873	0.6233	0.4174	0.1135
PGRF-CIRF	0.6314	0.3949	***0.6167***	***0.4472***	0.1527
SVM_{Choi}	0.5980	0.3869	0.6335	0.3681	0.0175
PGRF-SVM_{Choi}	***0.6355***	***0.4030***	0.6336	0.4370	***0.1607***†
SVM_{Qu}	0.5379	0.2103	10.6114	0.3686	0.0228
PGRF-SVM_{Qu}	0.4951	0.1989	10.2646	0.3692	0.0236

Table 4.8: Classification results on CME dataset.

Figure 4.11 plots the most important CME and CNN features. These are calculated from the CIRF model according to the conditional permutation importance, which adjusts for correlations between features [61]. In this way, a fair assessment as to which feature are most important can be investigated.

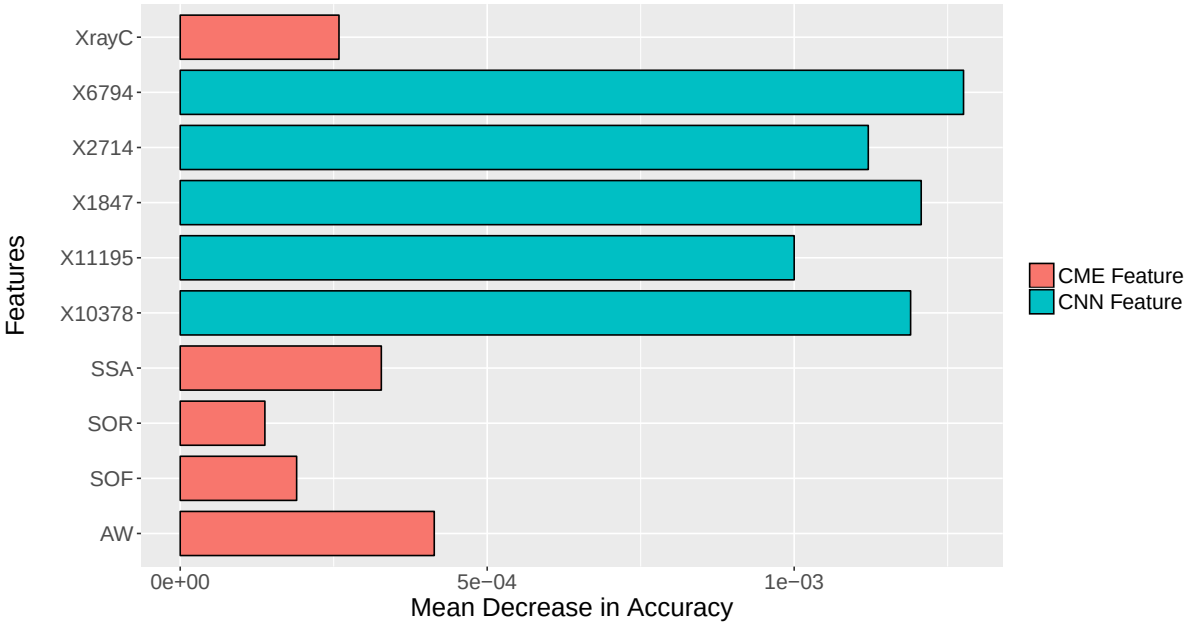


Figure 4.11: Top five most important CME and CNN features according to the PGRF-CIRF.

Clearly, by analyzing the top 5 from each feature type, the CNN features are considered more useful for making predictions, further adding to the evidence of their value. On an additional note, although the PGRF-CIRF does not yield the best performance in

all the metrics (only 2/5), the supplemental benefit to reliably analyzing the importance of such features demonstrates the advantages using RF based approaches in CME classification, as seen in other works [42], as opposed to SVM based ones, since these typically have no feature importance scheme.

Overall, these results show improvement in including these deep learning features into initial CME classification. Given the effort to include these is minimal (since the only real computational effort is the selection of the feature maps), it is imperative to include this type of information to make timely and accurate predictions to this difficult classification problem.

4.5 Discussion

In this work, a modification of the GRF model as well as an architecture for improving initial CME classification is proposed. In the former, instead of using the Gini importance to derive the penalties for each feature in a RF, using those based on the permutation importance scheme for high-dimensional data helps correct for much of the bias in RFs in high-dimensional settings. Simulation studies demonstrated the opportunity for more reliable feature selection with the PGRF counterpart. Real data studies showed that PGRFs can offer similar performance while being much faster than the GRF. This feature selection method can be used to help include relevant deep learning features alongside traditional CME information for making initial classifications, an opportunity that has not been explored in space weather studies. As shown in this work, integrating PGRF selected feature maps from a pre-trained CNN can offer better predictions, sometimes with significant difference, which is vital since this data can lead to preparations being made on Earth days in advance.

Many future work opportunities exist. First, only one set of parameter settings are implemented for the PGRF. Investigating the performance from a variety of parameters, such as γ and α, can yield better performance. Second, only two classifiers are executed on the CME dataset. Examining the combination of deep learning features with other classifiers, or within multi-step approaches, such as the two-stage meta-learning frame-

work posited by Larkin [43], is worth exploring.

4.6 Conclusion

CMEs remain a constant threat to modern society due to their potential impact on technology. While accurate predictions can be made using data obtained closer to Earth, this leaves very little time to make preparations on Earth. Using data gleaned at the onset of a CME gives more lead time, but generally leads to suboptimal predictions. Hence, it is vital to explore ways of increasing predictive performance once a CME is detected. As shown here, one way to do this is to exploit a CME's image information. Convolving these images with a pre-trained CNN can offer relevant features for prediction by providing higher-level abstractions of the image. Since CME images are already being collected and can easily be introduced into pre-trained CNNs, this provides an effective way to increase the predictive power of models in initial CME classifications without sacrificing the time advantage. As computing power and the number CME of images being stored increases, the opportunities to utilize deep learning for mitigating the potential impacts of dangerous geomagnetic storms will continue to grow, thereby, saving business entities millions of dollars in damages and loss of opportunity.

4.7 References

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[2] Revolution Analytics and Steve Weston. *foreach: Provides foreach looping construct for R*, 2015. R package version 1.4.3. Retrieved from `https://CRAN.R-project.org/package=foreach` [accessed: 2017-06-09].

[3] Ben Athiwaratkun and Keegan Kang. Feature representation in convolutional neural networks. *arXiv preprint arXiv:1507.02313*, 2015.

[4] Simon Bernard, Sébastien Adam, and Laurent Heutte. Using random forests for handwritten digit recognition. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 1043–1047. IEEE, 2007.

[5] Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in Knowledge Discovery and Data Mining*, pages 3–12. Springer, 2004.

[6] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] HV Cane and IG Richardson. Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, 108(A4), 2003.

[8] Hwann-Tzong Chen, Tyng-Luh Liu, and Chiou-Shann Fuh. Segmenting highly articulated video objects with weak-prior random forests. In *European Conference on Computer Vision*, pages 373–385. Springer, 2006.

[9] Tianqi Chen, Qiang Kou, and Tong He. *mxnet: MXNet*, 2015. R package version version 0.9.4. Retrieved from `https://github.com/dmlc/mxnet/tree/master/R-package` [accessed: 2017-06-09].

[10] Seonghwan Choi, Yong-Jae Moon, Ngo Anh Vien, and Young-Deuk Park. Application of support vector machine to the prediction of geo-effective halo CMEs. *J. Korean Astron. Soc*, 45:31–38, 2012.

[11] Yu Dai, Wei-guo Zong, and Yu-hua Tang. A quantitative research on the classification of coronal mass ejections. *Chinese Astronomy and Astrophysics*, 26(2):183–188, 2002.

[12] Houtao Deng. Guided random forest in the RRF package. *arXiv preprint arXiv:1306.0237*, 2013.

[13] Houtao Deng and George Runger. Feature selection via regularized trees. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

[14] Houtao Deng and George Runger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483–3489, 2013.

[15] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3, 2006.

[16] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014.

[17] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[18] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-76. Retrieved from `https://CRAN.R-project.org/package=caret` [accessed: 2017-06-09].

[19] G. Riddick et al. Predicting in vitro drug sensitivity using random forests. *Bioinformatics*, 27(2):220–224, 2011.

[20] Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, 2008.

[21] Irving John Good. Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *The British Journal for the Philosophy of Science*, 19(2):123–143, 1968.

[22] N Gopalswamy, S Yashiro, G Michalek, G Stenborg, A Vourlidas, S Freeland, and R Howard. The SOHO/LASCO CME catalog. *Earth, Moon, and Planets*, 104(1-4):295–313, 2009.

[23] Isabelle Guyon, Steve R Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In *NIPS*, volume 4, pages 545–552, 2004.

[24] Trevor Hastie and Junyang Qian. glmnet vignette, 2014. Retrieved from `http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf` [accessed: 2017-06-06].

[25] Douglas M Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[27] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[28] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2005.

[29] RA Howard, NR Sheeley, MJ Koomen, and DJ Michels. Coronal mass ejections: 1979–1981. *Journal of Geophysical Research: Space Physics*, 90(A9):8173–8191, 1985.

[30] Tim Howard. *Coronal Mass Ejections: An Introduction*, volume 376. Springer Science & Business Media, 2011.

[31] Jin Huang and Charles X Ling. Using AUC and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 17(3):299–310, 2005.

[32] Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, pages 1–31, 2015.

[33] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[34] Andrej Karpathy. Convolutional neural networks (CNNs / ConvNets). Retrieved from `http://cs231n.github.io/convolutional-networks/#fc` [accessed: 2017-06-06].

[35] Andrej Karpathy. Transfer learning. Retrieved from `http://cs231n.github.io/transfer-learning/` [accessed: 2017-06-06].

[36] Andrej Karpathy. What I learned from competing against a ConvNet on ImageNet. Retrieved from `http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/` [accessed: 2017-06-06].

[37] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

[38] R-S Kim, Y-J Moon, N Gopalswamy, Y-D Park, and Y-H Kim. Two-step forecast of geomagnetic storm using coronal mass ejection and solar wind condition. *Space Weather*, 12(4):246–256, 2014.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[40] Max Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[41] Zhenzhong Lan, Yi Zhu, and Alexander G Hauptmann. Deep local video feature for action recognition. *arXiv preprint arXiv:1701.07368*, 2017.

[42] Taylor Larkin. A tree ensemble for classifying geoeffective coronal mass ejections, 2016. Working paper.

[43] Taylor Larkin. A two-stage meta-learning framework for predicting geomagnetic storms, 2016. Working paper.

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

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

[46] M. Lichman. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2013. Retrieved from `http://archive.ics.uci.edu/ml` [accessed: 2017-05-05].

[47] Sanya Liu, Zhi Liu, Jianwen Sun, and Lin Liu. Application of synergetic neural network in online writeprint identification. *International Journal of Digital Content Technology and its Applications*, 5(3):126–135, 2011.

[48] Lin Ma, Zhuo Chen, Long Xu, and Yihua Yan. Multimodal deep learning for solar radio burst classification. *Pattern Recognition*, 61:573–582, 2017.

[49] Guy Nimrod, András Szilágyi, Christina Leslie, and Nir Ben-Tal. Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. *Journal of Molecular Biology*, 387(4):1040–1053, 2009.

[50] National Oceanic and Atmospheric Administration. Index of /pub/warehouse. Retrieved from `ftp://ftp.swpc.noaa.gov/pub/warehouse` [accessed: 2015-04-15].

[51] Antti Pulkkinen, Sture Lindahl, Ari Viljanen, and Risto Pirjola. Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system. *Space Weather*, 3(8), 2005.

[52] Ming Qu, Frank Y Shih, Ju Jing, and Haimin Wang. Automatic detection and classification of coronal mass ejections. *Solar Physics*, 237(2):419–431, 2006.

[53] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. Retrieved from `https://www.R-project.org/` [accessed: 2017-06-09].

[54] Vijay Raghavan, Peter Bollmann, and Gwang S Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

[55] IG Richardson and HV Cane. Near-Earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009): Catalog and summary of properties. *Solar Physics*, 264(1):189–237, 2010.

[56] E Robbrecht and D Berghmans. Automated recognition of coronal mass ejections (CMEs) in near-real-time data. *Astronomy & Astrophysics*, 425(3):1097–1106, 2004.

[57] Guzman Santafe, Iñaki Inza, and Jose A Lozano. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508, 2015.

[58] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[59] SOHO: Solar and Heliospheric Observatory. Data/archive. Retrieved from `https://sohodata.nascom.nasa.gov/cgi-bin/data_query` [accessed: 2017-06-09].

[60] Nandita Srivastava and P Venkatakrishnan. Solar and interplanetary sources of major geomagnetic storms during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, 109(A10), 2004.

[61] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

[62] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

[63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[64] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[65] Vladimir Naumovich Vapnik. *Statistical Learning Theory*, volume 1. Wiley New York, 1998.

[66] Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011.

[67] Martin Vincent and Niels Richard Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786, 2014.

[68] Ian H Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[69] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[70] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[71] Guanwen Zhang, Jien Kato, Yu Wang, and Kenji Mase. How to initialize the CNN for small datasets: Extracting discriminative filters from pre-trained model. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 479–483. IEEE, 2015.

[72] J Zhang, KP Dere, RA Howard, and V Bothmer. Identification of solar sources of major geomagnetic storms between 1996 and 2000. *The Astrophysical Journal*, 582(1):520, 2003.

[73] Ling Zhang, Jian-qin Yin, Jia-ben Lin, Xiao-fan Wang, and Juan Guo. Detection of coronal mass ejections using AdaBoost on grayscale statistic features. *New Astronomy*, 48:49–57, 2016.

5

OVERALL CONCLUSION

In each of these works, the main focus is the development and application of novel ensemble models to make predictions regarding the severity of geomagnetic storms based on CME data. In practice, these models can be operationalized by entities such as the Space Weather Prediction Center, which is maintained by NOAA. The predictions can translate into better alerts that NOAA can disseminate to other government agencies and businesses where this type of space weather is a concern. Similar to other natural disasters, geomagnetic storm strength relates to alert level (as seen in Chapter 2). With these warnings, the proper preparations can be made, whether it means shutting down power grids or suspending air traffic. More information regarding the types of warnings and alerts available can be found here: `http://www.swpc.noaa.gov/`.

Given our dependency on telecommunications and commercial satellites, any disruption in these services could cost millions of dollars for corporations and government agencies worldwide. CMEs, being the primary driver of severe geomagnetic storms, remain a constant threat to modern society due to their potential impact on technology. Fortunately, satellite data allows empirical studies to be done on these events, both for inference and prediction tasks. Hence, it is possible to construct comprehensive datasets to benchmark new algorithms and models (descriptive analytics). Not only can ensembles provide superior predictive performance, but they can also provide model based insights (predictive analytics). By using flexible approaches for variable selection such as those based on RFs, reliable inspection as to what CME information is most vital to analyze when predicting strong geomagnetic disturbances, especially as new satellites collect new data (Chapter 2). In terms of prediction, accurate forecasts can be made using data obtained closer to Earth, but at the sacrifice of lead time. Using data gleaned at the onset of a CME gives more lead time but generally leads to suboptimal forecasts. Hence,

it is imperative to explore ways establishing a balance between this trade-off (prescriptive analytics). This can be done effectively by establishing a two-stage approach (Chapter 3). In addition, initial CME classification can be improved through use of deep learning and careful feature selection (Chapter 4). Future work consists of conducting an extensive study on quantifying costs to telecommunications and the cost-savings for implementing the proposed two-stage framework in practice. Due to the potentially catastrophic consequences CMEs can have on the global business environment, studying such space weather events is an absolute necessity using the most advanced analytical tools from both machine and statistical learning literature.