10-1-2019 2:45 PM

Machine Learning Classification of Interplanetary Coronal Mass Ejections Using Satellite Accelerometers

Kelsey Doerksen, *The University of Western Ontario*

Supervisor: Dr. Kenneth McIsaac, *The University of Western Ontario*
Co-Supervisor: Dr. Jayshri Sabarinathan, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Electrical and Computer Engineering

Abstract

Space weather phenomena is a complex area of research as there are many different variables and signatures that are used to identify the occurrence of solar storms and Interplanetary Coronal Mass Ejections (ICMEs), with inconsistencies between databases and solar storm catalogues. The identification of space weather events is important from a satellite operation point of view, as strong geomagnetic storms can cause orbit perturbations to satellites in low-earth orbit. The Disturbance storm time (Dst) and the Planetary K-index (Kp) are common indices used to identify the occurrence of geomagnetic storms caused by ICMEs, among several other signatures that are not consistent with every storm. Moreover, specific instrumentation is needed for solar storm and space weather phenomena, which can be costly and technically difficult for small and nano-satellite applications. This thesis demonstrates the capability of a new signature for identification and characterization of ICMEs, through the use of satellite accelerometer data from the Gravity Recovery and Climate Experiment (GRACE) satellite, and machine learning techniques. Utilizing pre-existing satellite instrumentation, this research proposes the use of accelerometers for future space weather monitoring applications. Four binary classification algorithms have been explored: Random Forest, Support Vector Machine, Extremely Randomized Trees, and Logistic Regression. It is proposed that a binary classification model can differentiate between a solar storm caused by an ICME versus a period of quiet geomagnetic activity, using only the accelerometer data of a satellite. Of the four architectures, the tree-based machine learning models performed the best, with accuracy scores over 80%.

Keywords: Interplanetary Coronal Mass Ejection, Machine Learning, Space Weather, Random Forest, Extremely Randomized Trees, Logistic Regression, Support Vector Machine

Lay Summary

Space weather phenomena is a complex area of research. An eruption of energy on the surface of the Sun sends high-energy particles towards Earth, a signature of the beginning of an event known as an Interplanetary Coronal Mass Ejection (ICME). When these events reach the Earth's atmosphere, they result in geomagnetic storms, which physically alter the atmosphere around the Earth and the satellites orbiting within it. ICME and geomagnetic storm events are difficult to characterize, as there are many different variables and signatures that are used to identify them, with inconsistencies between databases and storm catalogues. The identification of space weather events is important from a satellite operation point of view, as strong geomagnetic storms can cause the orbit properties of a satellite to change unexpectedly, which could result in collisions with other spacecraft, or unwanted re-entry. The Disturbance Storm-Time (Dst) and the Planetary K-index (Kp) are common indices used to identify the occurrence of geomagnetic storms caused by ICMEs, among several other signatures that are not consistent with every storm. Moreover, specific instrumentation is needed for solar storm and space weather phenomena research, which can be costly and technically difficult for small and nano-satellite applications. This thesis demonstrates the capability of a new signature for identification and characterization of ICMEs, through the use of satellite accelerometer data from the Gravity Recovery and Climate Experiment (GRACE) satellite, and machine learning techniques. Utilizing pre-existing satellite instrumentation and extracting statistical information from what is physically measured by a satellite, this research proposes the use of accelerometers for future space weather monitoring applications. Four binary classification algorithms have been explored: Random Forest, Support Vector Machine, Extremely Randomized Trees, and Logistic Regression. It is proposed that a binary classification model can differentiate between a solar storm caused by an ICME versus a period of quiet geomagnetic activity, using only the accelerometer data of a satellite.

Acknowledgments

I would like to first thank my supervisors, Dr. Kenneth McIsaac and Dr. Jayshri Sabarinathan, for their support throughout my research work, and allowing me to pursue a topic outside of the scope of their usual Masters student researchers. I would also like to thank my supervisors and colleagues at l'Observatoire de Paris; Dr. Carine Briand, Dr. Florent Deleflie, Dr. Luc Sagnières, and Ali Sammuneh, for their support, guidance, and friendship during my internship work and beyond. I want to thank my research group, especially Matt Cross and Alexis Pascual, for always offering their help when I needed it. Thank you to my family, for supporting me throughout this long academic journey. Finally, I want to thank my partner, for their constant support, encouragement, and positivity throughout this journey, no matter how far away they were.

Contents

List of Figures

List of Tables

List of Appendices

List of Acronyms

3DP	3D Plasma and Energetic Particles Experiment
ACE	Advanced Composition Explorer
AMR	Area-to-Mass Ratio
AUC	Area under the ROC curve
CACTUS	Capteur Accelerometrique Triaxial Ultra Sensible
CHAMP	Challenging Minisatellite Payload
CIR	Co-rotating Interaction Regions
CME	Coronal Mass Ejection
DLR	Deutsche Forschungsanstalt für Luft und Raumfahrt
Dst	Disturbance storm time
ERT	Extremely Randomized Trees
ESA	European Space Agency
EUV	Extreme Ultraviolet
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GFZ	German Research Centre for Geosciences
GME	Goddard Medium Energy
GRACE	Gravity Recovery and Climate Experiment
GRACE-FO	Gravity Recovery and Climate Experiment Follow On
ICME	Interplanetary Coronal Mass Ejection
IMF	Interplanetary Magnetic Field
JPL	Jet Propulsion Lab
Kp	Planetary K-index
LR	Logistic Regression
MFI	Magnetic Field Investigation

MGS	Mars Global Surveyor
NASA	National Aeronautics and Space Administration
OGO	Orbiting Geophysical Observatory
QDA	Quadratic Discriminant Analysis
RF	Random Forest
ROC	Receiver Operating Characteristic
SEE	Solar EUV Experiment
SETA	Satellite Electrostatic Triaxial Accelerometer
SNR	Signal to Noise Ratio
SOHO	Solar and Heliospheric Observatory
STAR	Space Three-axis Accelerometer for Research
SuperSTAR	Super Space Three-axis Accelerometer for Research
SVM	Support Vector Machine
SWE	Solar Wind Experiment
SWEPAM	Solar Wind Electron, Proton, and Alpha Monitor
SWICS	Speeds, Composition, and Charge States
TGCM	Thermospheric General Circulation Model
TIMED	Thermosphere Ionosphere Mesosphere Energetics and Dynamics
TN	True Negative
TP	True Positive
UTC	Universal Time Coordinated
UV	Ultraviolet
WIND	Global Geospace Science

Chapter 1

Introduction

1.1 Motivation

The motivation for this work is to develop a predictive method of identifying ICMEs using pre-existing satellite instrumentation with minimal data pre-processing. Specifically, the use of the Level1B data from the on-board SuperSTAR accelerometer of the GRACE-A satellite to identify ICMEs. It is proposed that a machine learning classifier can be used with the accelerometer data of this spacecraft to identify ICME-related events. Similar algorithms could also be implemented on other spacecraft at varying altitudes within the thermosphere to better:

- Understand the atmosphere's response to ICMEs (i.e., are more ICME's identified at different altitudes);
- Provide a space weather detection protocol for any spacecraft;
- Increase the overall knowledge of space weather phenomena;
- Increase the understanding of spacecraft interaction with solar storms in our atmosphere and;
- Present a new characteristic signature to identify the passage of ICMEs.

The motivation for this work is to be able to use a binary classifier to distinguish between satellite behaviour during a solar storm caused by an ICME versus satellite behaviour during periods of quiet geomagnetic activity. This work uses data that has gone through little processing compared to the past research work published using the GRACE thermospheric density estimate datasets to study space weather as described in Chapter 2. The Level1B accelerometer data is used to extract information (in this case, identification of solar storm occurrence) about space weather phenomena in a new way, opening the door to the additional utilization of satellite attitude determination sensors to study space weather.

The inspiration behind the beginning of this research came from a joint-endeavour between l'Observatoire de Paris and the University of Western Ontario performed by the author and colleagues, to study the effects of solar flare events on the orbiting spacecraft within Earth's atmosphere. The GRACE and CHAMP satellites were chosen as test subjects because they have similar lifetime data available, there has been already defined accelerometer-derived thermospheric density datasets from numerous sources, and past literature has used these spacecraft in observing the thermosphere's density changes caused by solar events. It is key to study the effects of solar flares in particular because there has been little research performed in this field as to the effects of flares in isolation, especially with respect to the orbital ephemeris of spacecraft. X-class flare events were considered when observing the density fluctuations due to solar flares, as they are the strongest class. The study included performing simulations of the expected satellite decay due to atmospheric density increase with results shown in Figure 1.1. These results created a further motivation to understand the varying effects of space weather phenomena on orbiting satellites, as reduction in mission lifetime and in-orbit satellite collision are all very possible realities to spacecraft caused by solar events.

S. Krauss et al. [5] detailed a statistical analysis of the thermosphere and geomagnetic response to ICMEs observed by the ACE and GRACE spacecraft, which found that the majority

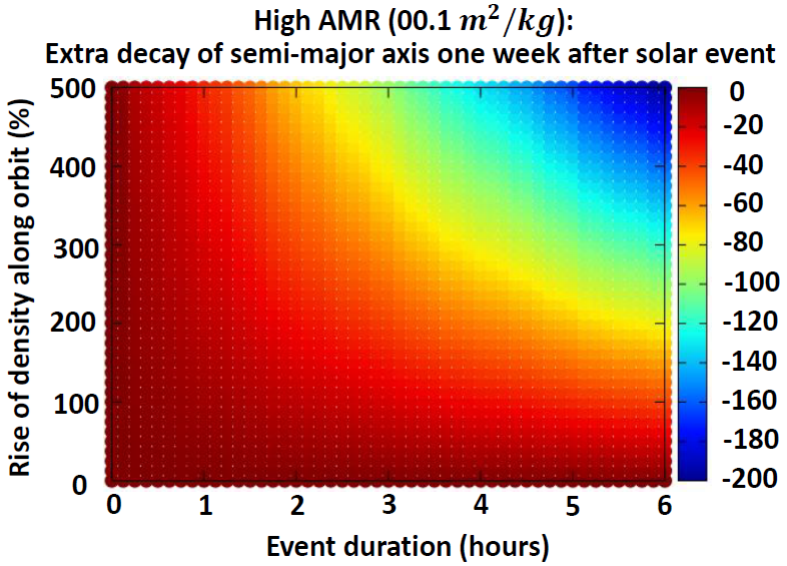

Figure 1.1: Simulated satellite semi-major axis decay due to density increase from solar event. (AMR = area-to-mass ratio of satellite, colour bar is satellite semi-major axis decay in meters)

of ICMEs studied caused a significant increase to the Earth's thermosphere neutral density levels. Their results provide support to the notion that the GRACE accelerometer data, from which the thermospheric density levels were derived from, can be used in a machine learning model to identify periods of increased geomagnetic activity caused by ICMEs, which are accompanied by these density increases. Additionally, Sutton et al. [1] examined the neutral density response to the large solar flare events during the 2003 "Halloween storm", again using the accelerometer-derived densities of GRACE and CHAMP to study the thermosphere's perturbed behaviour. Figure 1.2 shows the total atmospheric mass density changes using the accelerometer-derived density datasets to observe the increase in density caused by the 2003 Halloween solar storm. The left shows GRACE densities normalized to 490 km, at local times (bottom) 04:00:00 and (top) 16:00:00. The right of the figure shows CHAMP densities normalized to 400 km, at local times (bottom) 01:20:00 and (top) 13:20:00. EUV fluxes from the SEE instrument on TIMED orbiter are superimposed in the top plots. The Kp and North Polar Cap Index (commonly used to quantify magnetic activity) are superimposed in the left and right bottom plots, respectively.

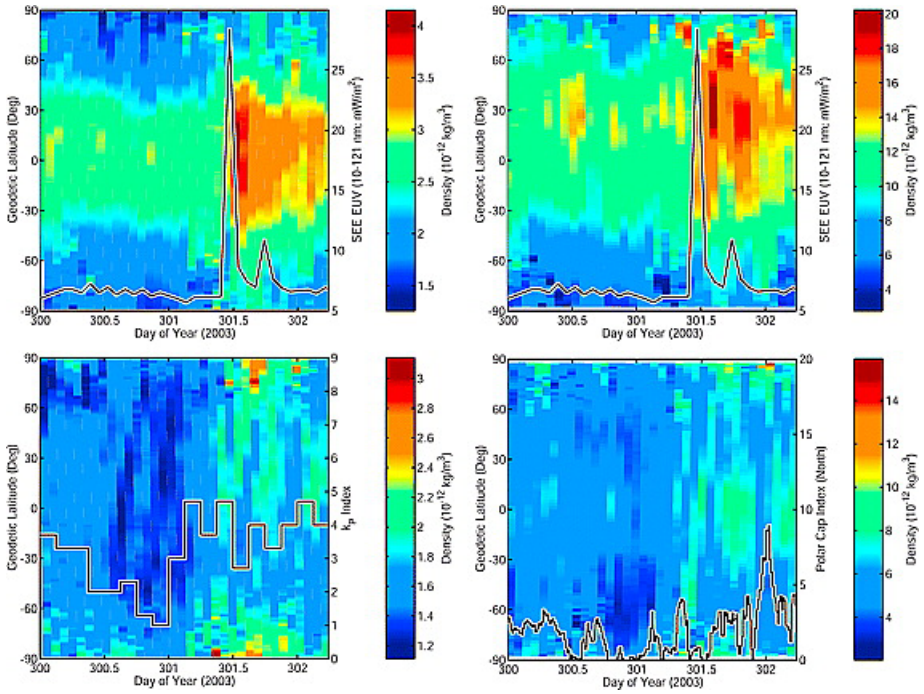

Figure 1.2: Latitude versus time depictions of total mass density measured during 27-28 October (day 300-301), 2003. [1].

Figure 1.2 clearly shows that there is an increase in density from the change in the colourmap over the duration of the event, information which has been derived from the accelerometer of the GRACE and CHAMP satellites. The question that was then posed to be answered with this thesis work is, *"If the atmospehere's density, which changes and fluctuates due to solar storms, can be derived using the on-board accelerometer of GRACE, is it possible to see similar changes and fluctuations in the raw, unprocessed data, that could be used for space weather monitoring applications?"* In particular, ICMEs are of interest in this research as there is a multitude of previous literature supporting their atmospheric changes due to solar storm events measured by GRACE and CHAMP accelerometers as opposed to the more limited amount from solar flare events, and they are the predominant drivers of intense geomagnetic storms, which affect both satellite instrumentation and orbitography.

The applications of an ICME classification system using accelerometer data are especially relevant to satellite operators. An automatic detection system to identify ICMEs that have per-

turbed a spacecraft can provide information that can be used to make informed decisions for in-orbit attitude adjustment. From a space weather research perspective, utilizing accelerometer data is an additional piece of information to characterize ICME events, without the need for additional equipment to do so. For SmallSat and CubeSat operators, using a lower-resolution accelerometer to study space weather events could be a less expensive alternative to imaging and radiation-measurement-based instruments.

1.2 Tasks and Thesis Contribution

This research explores the use of unprocessed Level1B satellite accelerometer data to identify geomagnetic storms caused by ICMEs. It is proposed that a spacecraft's on-board attitude determination sensors can be used for more than just orbit control and adjustments; to be proven via the use of GRACE-A satellite. Specifically, the accelerometer data from the on-board SuperSTAR instrument of GRACE-A can be used to develop a machine learning classification model of the ICME events that occur over its lifetime. ICME storms can be automatically classified from periods of quiet geomagnetic activity by extracting useful, important features from the data and using Random Forest, Extremely Randomized Trees, Support Vector Machine, and Logistic Regression machine learning architectures.

Using the satellite accelerometer data prior to processing with thermospheric and drag modelling is a potentially new method of utilizing machine learning for space weather applications, and gives insight as to how what is physically felt by the spacecraft can be translated into information for the space weather community.

1.3 Thesis Outline

The remainder of the thesis is organized as follows: Chapter 2 contains a literature review detailing the GRACE satellites, a general overview of space weather phenomena and the atmospheric and spacecraft response to such events, previous use of satellite accelerometers for space weather monitoring, RF, ERT, SVM, and LR machine learning techniques, and previous work done in the field of machine learning for space weather applications. Chapter 3 details previous work related to classification of accelerometer data, and the benchmark model used for reference. Chapter 4 details the methodology of the thesis, including data acquisition, class balancing, any pre-processing needed prior to machine learning, feature vector extraction, and the four proposed machine learning architectures used with the dataset. Chapter 5 presents the performance of the classifiers of each architecture, Chapter 6 discussion, and finally Chapter 7 the conclusion of the thesis work and proposed future direction of the research.

Chapter 2

Literature Review

2.1 Gravity Recovery and Climate Experiment

The GRACE mission was a joint partnership between NASA in the United States and the DLR in Germany, implemented under the NASA Earth System Science Pathfinder Program. The mission consisted of two twin satellites, following approximately 220 km from one another in a polar orbit 500 km above Earth, which allowed for a 30 second separation of the satellites in the same orbital plane (Figure 2.1). The mission was launched on March 17th, 2002 and lasted until October 2017, originally intended to last only five years. The GRACE mission was developed to accurately map the variations in the Earth's gravity field, by making accurate measurements of the distance between the two satellites using GPS and a microwave ranging system [15]. The primary objective of the GRACE mission was to map the global gravity field with a spatial resolution of 400-40,000 km every thirty days, using, among other instrumentation, the SuperSTAR accelerometer [2], [16].

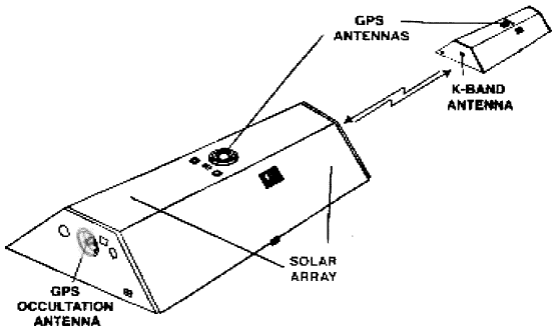

Figure 2.1: GRACE satellite [2]

2.1.1 The STAR and SuperStar Accelerometers

The STAR accelerometer integrated into the center of mass of the CHAMP satellite, performed the measurements of the non-gravitational accelerations, for purposes of recovering the Earth's gravity field [17]. There is considerable heritage in the design of the GRACE satellite as compared to the CHAMP mission, including this instrument. Figure 2.2 depicts the unit before integration into the spacecraft. STAR is a six-axis accelerometer that provides a measurement range of $\pm\ 10^{-4} m/s^2$ and exhibits a resolution of better than $3x10^{-9}\ m/s^2$ for the y- and z-axes and $3x10^{-8}\ m/s^2$ for the x-axis within the measurement bandwidth from 10e-4 to 10e-1 Hz [3]. Orbitography was performed with the accelerometer, with an accuracy of a few millimetres that measured the air drag, solar and Earth radiation pressures, and the attitude manoeuvre effects.

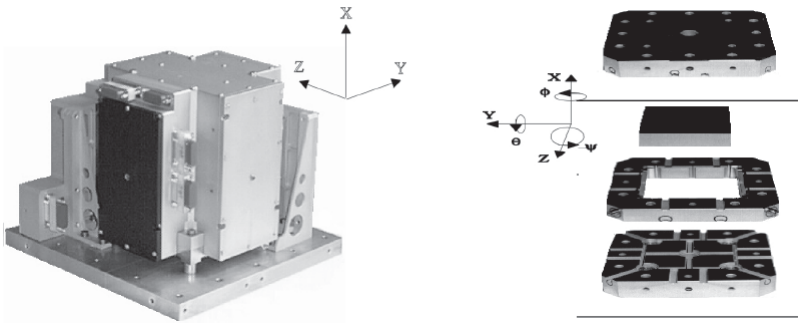

Figure 2.2: STAR sensor unit before integration in the CHAMP satellite (right) and its internal core (left) [3].

The SuperSTAR accelerometer aboard GRACE was almost identical to the STAR accelerometer, with an increased resolution of $10e-10\ ms^{-2}\ \sqrt{Hz}$. The increased resolution allowed for a range to 175 μm. The accelerometer measured electrostatic forces proportional to the acceleration of the satellite that are caused the non-gravitational forces acting on the spacecraft [18].

2.2 Space Weather

Space weather is the term used to describe the dynamic conditions in the Earth's outer space environment. It includes any and all conditions on the Sun, in the solar wind, in near-Earth space and in the upper atmosphere that can affect space-borne and ground-based technological systems [19]. Examples of space weather events include solar flares, Coronal Mass Ejections (CME), and ICMEs, although this thesis work focuses solely on ICMEs and the geomagnetic disturbances (storms) related to them.

2.2.1 Geomagnetic Storms

A geomagnetic storm is a major disturbance of the Earth's magnetosphere, occurring when there is an exchange of energy from solar wind into the space environment surrounding Earth [20]. The largest storms result from CMEs, in which billions of tons of plasma from the Sun arrive at Earth. Geomagnetic storms cause changes in the currents, plasmas, and magnetic fields in the Earth's magnetosphere, and are a result from the variations in the solar wind energy.

Geomagnetic Storm Indices

There are many signatures to identifying a geomagnetic storm, including plasma, solar wind, and magnetic field measurements. The Dst and Kp are the two key indices used throughout this

research to identify both periods of quiet and disturbed geomagnetic activity, and are detailed below.

The Dst is an index of magnetic activity derived from near-equatorial geomagnetic observatories that measure the intensity of the equatorial electrojet, an intense electric current which occurs in the lower ionosphere [21]. The Dst index is a measure of geomagnetic activity in nanoteslas, used to assess the severity of magnetic storms [22]. At the onset of a magnetic storm, the Dst shows a sudden rise, followed by a sharp decrease. In general, a Dst value of -50 and below indicates a significant geomagnetic disturbance.

The Kp index quantifies disturbances in the horizontal component of Earth's magnetic field with an integer in the range 0-9 with 1 being calm, 5 or more indicating a geomagnetic storm, and 9 indicating an intense storm [23]. The Kp index is a descriptor of the disturbance of the Earth's magnetic field caused by solar wind. The faster the solar wind blows, the greater the turbulence, the larger the Kp value.

2.2.2 Interplanetary Coronal Mass Ejections

Interplanetary Coronal Mass Ejections are solar wind structures that are generally believed to be the counterparts of Coronal Mass Ejections at the Sun [24]. From a perspective of geomagnetic disturbance level, ICMEs differ from flares mainly in their relative strength and longevity. ICMEs take hours from start to finish, and can be felt on Earth days after, whereas solar flares can last on the order of minutes and the affects on Earth are felt almost instantaneously. Figure 2.3 depicts a schematic of an ICME driving a shock ahead of it. A sheath of turbulent, ambient solar wind plasma that is compressed and heated separates the shock and the ICME.

The passage of an ICME can be indicated by various signatures; the presence of abnormally low solar wind proton temperatures is observed in most events, and they are principally identified based on solar wind plasma and magnetic field observations. Unusually low solar wind

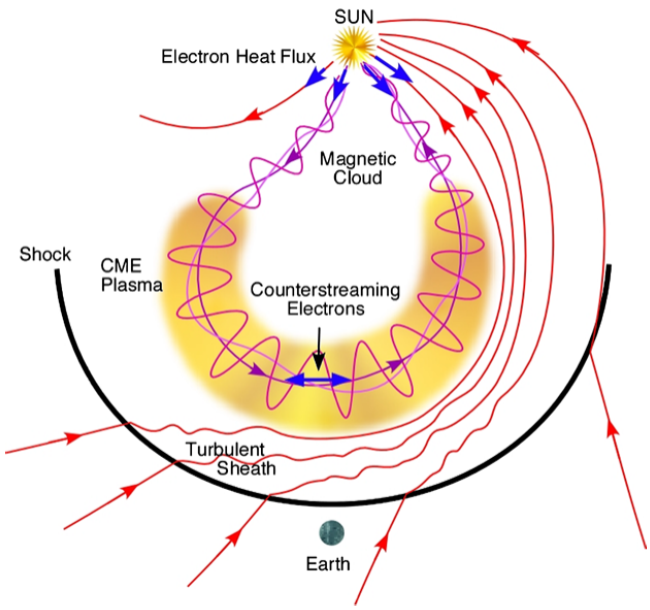

Figure 2.3: Schematic of an ICME and upstream shocks [4]

ion charge states are extremely rare but significant signatures, and have been found in very few ICMEs thus far in the field of solar wind monitoring as discussed by Schwenn, Rosenbauer and Muehlhauser [25], Gosling et al. [26], Zwickl et al. [27],Yermolaev et al. [28], Burlaga et al. [29], and Skoug et al. [30].

2.2.3 List of Richardson-Cane Interplanetary Coronal Mass Ejections from 1996-2019

The Richardson-Cane List of Interplanetary Coronal Mass Ejections from 1996-2019 was the sole database of ICMEs referenced for this thesis work. Their database was chosen because of their detailed methodology to identifying ICMEs, in which their "philosophy for identification was that ideally as many different signatures as possible should be examined when identifying ICMEs" [24]. Their database uses the following information from satellite instrumentation for increased precision in defining the ICME event and its start and end times: Near-Earth solar wind plasma and magnetic field data from IMP 8, WIND, and the ACE spacecraft, and the

OMNI near-Earth database at resolutions of 1 minute - 1 hour, half-hour solar wind composition and charge state observations made by the SWICS instrument on ACE, the solar wind electron pitch-angle from the SWEPAM/ACE spacecraft, energetic particle data from the IMP 8, ACE, and the SOHO spacecraft, and the galactic cosmic ray intensity at Earth using GME instrument on IMP 8. As of their revision on August 16th, 2019, the list encompasses 515 ICMEs from May 27th, 1996 to May 2nd, 2019.

2.3 Earth's Atmosphere

The Earth's atmosphere is divided into five layers: the troposphere (0-12km), stratosphere (12-50km), mesosphere (50-80km), thermosphere (80-700km), and exosphere (700-10,000km). The GRACE satellite mission referred to heavily in this study was within the Earth's thermosphere region, and therefore this will be the only atmospheric layer discussed in detail. The thermosphere region is the region of the atmosphere where the composition shifts from being largely composed of oxygen, to mostly atomic oxygen. Many low-earth orbit satellites carry most of their activities in the thermosphere [31], making it a particularly interesting region of the atmopshere to study, as many spacecraft that are susceptible to space weather events orbit within it.

2.3.1 Atmospheric Response to Solar Events

The reaction of the Earth's thermosphere to changing solar and geomagnetic conditions occurs through a series of complex and interdependent processes. In response to energy and momentum deposition at high latitudes in the form of Joule heating, particle precipitation, and electric fields that drive neutral winds via ion-neutral collisions, a global circulation system is set up to redistribute mass, momentum, and energy [32]. It is well known that the neutral-density of the Earth's atmosphere increases due to it's interaction with solar eruptive events. Bruinsma

et al. [32] noted density increases on the order of 300-800% as measured by GRACE and CHAMP during the November 20-21 2003 solar and geomagnetic storms.

2.3.2 Satellite Response to Solar Events

In the 1960s, much of the knowledge of magnetic storm response of the thermosphere was derived from the orbital decay of satellites [32]. Krauss et al.'s work, "Multiple satellite analysis of the Earth's thermosphere and interplanetary magnetic field variations due to ICME/CIR events during 2003-2015" [5] details the response of the GRACE and CHAMP spacecraft to ICME and CIR (co-rotating interaction regions) events over the course of their study period. In particular, they detail the magnitude of orbit decay of GRACE and CHAMP in relation to event strength, represented as the magnitude of the north-south interplanetary magnetic field (IMF) vector B_z, which is a commonly used index to indicate geomagnetic storm strength. Figure 2.4 shows the orbital decays versus orbital altitude and event strength in terms of B_z [nT] measurements for CHAMP and GRACE. The conclusions of Krauss' work states that extreme ICME-induced geomagnetic storms may lead to decay rates of up to 40-50m and 90-120m, for GRACE and CHAMP respectively.

2.4 Space Weather Monitoring

Space weather events are observed using both ground-based and space-based sensors and imaging systems. Telescopes are used to detect visible and ultraviolet light, as well as gamma and x-rays. Receivers and transmitters are used in the detection of radio shock waves caused by the interaction of CMEs and solar wind. Particle detectors are used to count ion and electron changes and magnetometers record changes in the Earth's magnetic field during geomagnetic storms. Auroral patterns above Earth are observed using UV and visible cameras. The following subsection will detail satellite-specific space weather monitoring capabilities, both with

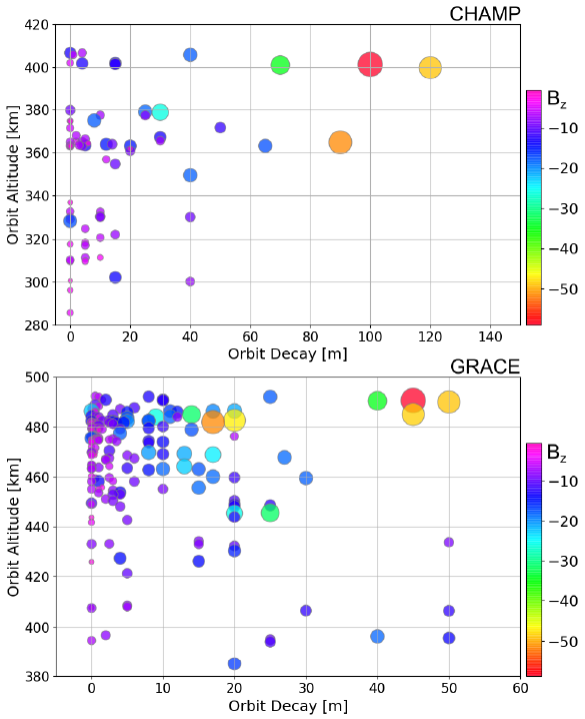

Figure 2.4: Orbit decays versus orbital altitude and event strength in terms of B_z [nT] measurements for the CHAMP and GRACE spacecraft [5].

and without the use of machine learning.

2.4.1 Previous Work

Space weather identification and monitoring has been done for many years. Complete conclusions of statistical studies on ICMEs can be found in Chi et al. [33], Nieves-Chinchilla et al. [34], Mitsakou and Moussas [35], and Kiplua et al. [36]. Lepping et al. [37] proposed an automatic magnetic cloud detection method based on empirical thresholds, and Ojeda-Gonzalez et al. [38] proposed an alternative automatic magnetic cloud identification method based on the computation of a Spatio-temporal Entropy. Automated short-term prediction of solar flares using solar imaging has been done by Colak & Qahwaji [39], and Karimabadi et al. [40] developed a data-mining method called MineTool-TS used to provide a classification of data intervals containing flux transfer events or lack thereof. Zurbuchen and Richardson [4] de-

tail criteria generally used to detect ICMEs manually. Neutral mass spectrometers on-board the OGO, Esro-4, AEROS, and Atmosphere Explorer satellites provided information on the atmosphere's neutral composition variations during magnetic storms detailed by Bruinsma et al. [32] and Prölss [41].

2.4.2 Instrumentation for Interplanetary Coronal Mass Ejection Identification

As mentioned, the STAR and SuperSTAR accelerometers have been used to study the effects of solar eruptive events on the Earth's thermosphere, through the use of atmospheric-density derived data using drag and thermospheric modelling. Most recent derivativations of such data can be found in Mehta et al. [42]. Bruinsma et al. [32], Krauss et al. [43], [5], Xiong et al. [44], Liu [45], and Sutton [1] are just some of the many works that have been done on observing the atmosphere's response to solar events utilizing the derived density data of the SuperSTAR and STAR accelerometers aboard CHAMP and GRACE. Another exmaple is the SWICS instrument aboard ACE. SWICS was a spectrometer optimized for measurements of the chemical and isotropic composition of solar and interstellar matter. It was designed to determine the chemical and ionic-charge composition of the solar wind (a component of an ICME), the temperatures and mean speeds of all major solar-wind ions, and to resolve H and He isotopes of both solar and interstellar sources [46]. Additionally, the GME instrument onboard the IMP 8 spacecraft was used to measure the galactic cosmic ray intensity at earth, and can be used to identify the bidirectional field-aligned energetic ion flows that are often associated with the passage of ICMEs [24].

2.4.3 The Use of Satellite Accelerometers for Space Weather Monitoring

There are two main applications of ultra-sensitive accelerometers in the fields of Earth and Planetary observations. The first application is the measurement of the forces acting on the satellite, which can provide information on the atmospheric density variation and high-altitude winds. Examples of this application include the CASTOR D5B satellite with the CACTUS accelerometer [47], the triaxial accelerometer system of Marcos et al. [48], and missions aiming at the analysis of the Mars atmosphere; Keating et al.'s [49] work utilizing the MGS z-axis accelerometer to obtain the structure of Mars' upper atmosphere, and Bougher et al.'s [50] work examining the aero-braking environment experienced by the MGS spacecraft on Mars. The second main application is the global recovery of the gravity field of the Earth or another planets, the purpose of the GRACE and CHAMP satellites.

More specifically, satellite-borne accelerometers have contributed significantly to advancing the knowledge of the Earth's thermospheric response to geomagnetic storms. Berger and Barlier [51] and [52], studied the disturbed thermosphere and its asymmetrical structure during magnetic storms with the CACTUS accelerometer experiment. Forbes et al. [53] made comparisons between simulations from the National Center for Atmospheric Research TCGM, satellite electrostatic triaxial accelerometer measurements of neutral winds and total mass densities between 170 and 240 km, and mid-latitude Thomson scatter measurements of neutral exospheric temperatures, for the isolated magnetic disturbance occurring on March 22, 1979. Additionally, Forbes et al. [54] studied the lower atmosphere magnetic storm response utilizing the SETA experiment and Forbes et al. [55] used the CHAMP satellite STAR accelerometer to analyze the thermosphere density variations due to the April 15-24, 2002 solar events. Liu and Luhr [56] used the CHAMP STAR accelerometer to examine the strong disturbances of the upper thermospheric density due to three geomagnetic super-storms of 2003, noting density enhancements on the order of 400-800%. S. Bruinsman and R. Biancale [57] published

first results on the thermopsheric density changes due to CMEs as derived from the CHAMP satellite, data which has been studied further in-depth since. Accelerometer data is the ideal measurement in terms of relevance to satellite drag applications because the instrument can sense the drag force on satellites (although it must be remembered that the drag force is due to external forces in addition to atmospheric density including solar wind effects).

The European Space Agency (ESA) launched their SWARM mission on November 22nd, 2013 with the aim to investigate and understand the Earth's magnetic field. The mission is a constellation of three identical spacecraft, each equipped with an accelerometer used to measure the non-gravitational acceleration in its respective orbit and, in turn, provide information about air drag and solar wind [58]. Kodikara et al. [59], provided the first analysis of the SWARM accelerometer-derived thermospheric densities, and compared this to both physical and empirical model estimates.

Most recently, the GRACE-FO mission has been launched. The mission is a partnership between NASA and the GFZ, and is a successor to the original GRACE mission. GRACE-FO carries the successful work of GRACE while testing a new technology designed to improve the precision of its measurement system. Specifically, GRACE-FO will test an experimental instrument that uses laser light instead of microwaves to measure the separation distance between the two satellites. Such technology could improve the accuracy of the measurement by tenfold or more, which will lead to higher-resolution gravity missions.

In all of the above applications, the accelerometer data was post-processed to derive the atmospheric density within the region at which the satellite was orbiting, with respect to their applications to space weather monitoring.

2.4.4 Space Weather Monitoring with Machine Learning

In recent years, machine learning techniques specific to space weather applications have grown in popularity. Bobra & Ilonidis [60] utilized machine learning classification of solar active regions that produce a solar flare with or without a CME. Camporeale et al. [61] used a Gaussian Process to create a four-category probabilistic classification algorithm for solar wind. The algorithm was trained and tested successfully using OMNI data, with a median accuracy larger than 90% for all categories. Yang et al. [62] proposed an artificial intelligence technology called the simulated annealing genetic method for automatic detection of sunspots on full-disk solar images. The algorithm allowed for self-adaptive threshold derivation for detecting the umbra and penumbra of sunspots simultaneously. Wang et al. [63] used a support vector machine on magnetopause crossings measured by 23 different spacecraft to propose a new, three-dimensional mangetopause model. Camporeale et al. [61] provided an accurate method of solar wind classification into four classes using a Gaussian process. Miniere & Pincon [64] provided a neural-network based method to identify and classify electron and proton whistlers from in-situ data measurements. Most recently, Ngyuen et al. [65] used convolutional neural networks to estimate a similarity parameter combined with a post-processing method based on peak detection for automatic ICME detection from the WIND spacecraft in-situ measurements. Their algorithm uses 30 primary input variables extracted from the data provided by the MFI, SWE, and 3DP instruments. The algorithm was able to detect a maximum of 197 of the 232 ICMEs during the 2010-2015 period, including 90% of the ICMEs present in the lists of Nieves-Chinchilla et al. and Chi et al. [34], [33]. The minimal number of False Positives was 25 out of 158 predicted ICMEs, yielding a precision of 84% ± 2.6%. To test the robustness of their algorithm, the author removed some of the 30 input parameters to test the following criteria for identification of ICMEs: considering solely the magnetic field magnitude and components, considering the magnetic field, proton fluxes, and Beta, considering the proton fluxes only, and considering the densities of proton and alpha particles only. Precision-

recall curves were constructed for the four configurations in addition to using all parameters for ICME detection, and the average precision was found to be 0.593, 0.621, 0.486, 0.334, and 0.697 when considering the magnetic field components only, the magnetic field and proton fluxes, the proton fluxes only, the densities of the proton and alpha particles only, and all the data, respectively.

2.5 Machine Learning for Classification

Machine learning is the scientific study of algorithms and statistical models that computers use to perform a specific task effectively, and without the use of explicit instructions. Machine learning relies on patterns, and inferring these patterns to ultimately make a decision. A classifier is a prediction system that is trained on past behaviour, and uses this information to assign a class to the new, unseen data it is tested on. The two main types of classification models are binary and multi-classification. In binary classification, there are two classes for which the model must differentiate between, a positive, or "1" class, usually representing the instance of an event, and a negative, or "0" class, usually representing the lack thereof an event, dependent on the dataset and research question. A multi-classification machine learning model has greater than two classes to predict, and is therefore more complicated, and will not be touched on in this thesis work.

2.5.1 Features and Feature Sets

The past behaviour that a machine learning model is trained on consists of features of the dataset in question. A feature is a piece of information about the dataset used by a machine learning algorithm, and is also commonly referred to as an attribute. For example, for a time series dataset, in which there are multiple data points representing a single class event, the mean of the time series is a feature. Statistical features are very commonly used in machine

learning, and contain information about a dataset including the mean, maximum, minimum, range, variance, standard deviation, and more. Features are stored in what is known as a feature vector, which holds all extracted features from a data series, including the class label (for supervised machine learning). A feature set is the combination of all of the feature vectors that make up a dataset. For example, for a dataset consisting of 100, equal-length time series, taking 5 features per series, would yield a feature set of 100, 1×5 feature vectors representing the data.

2.5.2 Training, Testing, and Validation Sets

In machine learning classification, the dataset of interest must be split up into training and testing sets, such that the model has enough data to accurately and effectively learn and recognize the patterns and make predictions based on these patterns. Typically, the dataset is broken into 70-80% training, or learning sample, and 20-30% testing sample. The training sample refers to the observations that are used to build a classification model, while the testing sample refers to the observations used to determine the model's accuracy and performance in predictions. Although not always necessary, but good practice for providing evidence of robustness and accuracy of a classifier, is the division of the dataset into train, test, and validation subsets. A validation set is a 10-15% subset of the data, that is separated from the entire dataset prior to training, and is used to validate the performance of the final model after training, testing, and parameter tuning has been completed.

2.5.3 Supervised Learning

Supervised learning is a type of machine learning in which the user provides a set of features and includes a label that corresponds to the class of which these features represent. It requires that the algorithm's possible outputs are already known, and the data that is used to train the

model is labelled with the correct answer (correct class). The correctly labelled class will be included for each feature vector in the feature set of a supervised learning model. A feature, also known as an attribute, denotes a particular input variable used in a supervised learning problem. Candidate features, or selected features, denote all the input variables that are available for a given problem, and make up the feature set. The number of selected features represents the dimensionality of the input space to the model. The output refers to the target variable that defines the supervised learning problem, when it is categorical (i.e. is there an ICME storm, or is there no ICME storm in this thesis' case), it is referred to as a classification problem.

Decision Tree Algorithm

A decision tree can be thought of as a series of yes or no questions asked about the dataset of interest, which eventually leads to a prediction. Decision trees make classifications similar to humans, asking a sequence of queries about the data until a final decision is reached. They use a top-down approach, in which a root node creates binary (yes/no) splits until a certain criteria is met. The root (first, top) node represents the entire population of the data, or in some cases, a sample of the population, which gets divided into two or more homogeneous sets. The general idea of using a decision tree for classification is to create a training model, which learns decision rules inferred from the prior data, i.e. training data, to predict a class [6]. Figure 2.5 depicts an example of the basic architecture of a decision tree model.

A computer model of a decision tree has no prior knowledge and must learn everything about the problem from the data provided by the user. During training, the model is given historical data, relevant to the problem (in the case of this thesis work, features extracted from accelerometer data), and the model learns any relationships between the features of the dataset and the values, or class, the user wants to predict (the target). When the decision tree makes a prediction, it must be given the same data features and will give an estimate that is based on the

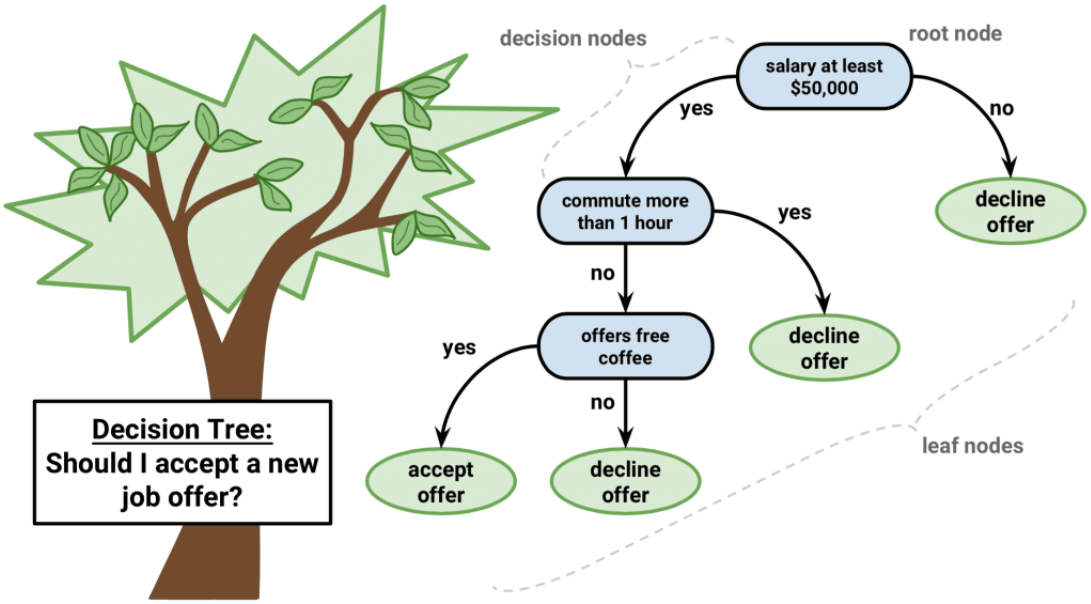

Figure 2.5: Decision Tree Example [6]

structures and relationships it has learned from training [66].

Within a decision tree, each decision node (also referred to as an ”internal” node) corresponds to a feature, and each leaf node corresponds to a class label. The psuedocode for the decision tree algorithm can be described as follows:

- Place the best feature of the dataset at the root of the tree;
- Split the training set into subsets, each subset contains data with the same value for a feature;
- Repeat steps on each subset until there are leaf nodes (class labels) present in all the branches of the tree.

The best feature of the dataset refers to the feature which could best classify the training examples on its own. The two most common measure to select the “best” feature are information gain and the gini index.

Information gain is the process used to estimate the information contained by each feature. It uses information theory to measure the randomness or uncertainty of a random variable, x, defined by its entropy. The feature with the largest information gain is then placed at the root of the decision tree, i.e., it is the "best" feature. The entropy of a feature is found from equation 2.1:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{2.1}$$

where H is the entropy and $p(x)$ is the probability of the class. By calculating the entropy of each feature, the information gain of each feature can be calculated. To illustrate this concept, consider the following decision tree construction example using the information gain criterion from [6].

For a dataset containing four columns of features, considered as predictors, A, B, C, and D, and one column of class labels (E), 0 or 1, a decision tree can be constructed converting continuous data (Figure 2.6) into categorical data.

	A	B	C	D	E
1	4.8	3.4	1.9	0.2	positive
2	5	3	1.6	0.2	positive
3	5	3.4	1.6	0.4	positive
4	5.2	3.5	1.5	0.2	positive
5	5.2	3.4	1.4	0.2	positive
6	4.7	3.2	1.6	0.2	positive
7	4.8	3.1	1.6	0.2	positive
8	5.4	3.4	1.5	0.4	positive
9	7	3.2	4.7	1.4	negative
10	6.4	3.2	4.5	1.5	negative
11	6.9	3.1	4.9	1.5	negative
12	5.5	2.3	4	1.3	negative
13	6.5	2.8	4.6	1.5	negative
14	5.7	2.8	4.5	1.3	negative
15	6.3	3.3	4.7	1.6	negative
16	4.9	2.4	3.3	1	negative

Figure 2.6: Information gain step one [6]

The following values are used to categorize each feature based on their values shown in Figure 2.6:

A	B	C	D
≥5	≥3.0	≥4.2	≥1.4
< 5	< 3.0	< 4.2	< 1.4

Table 2.1: Decision Tree Example: Feature Categorization

The information gain for each feature is found by first calculating the entropy of the target (in this case, it is a binary classification problem, the target is either 1 or 0, and there is 50/50 class balance, therefore the entropy of the target is 1), then calculating the entropy for each feature, and subtracting this entropy from the target entropy to yield the information gain. For example, the information gain for A is found as follows: 12 out of 16 records satisfy the criteria ≥5 (called criteria 1), and 4 out of 16 records satisfy the criteria <5 (called criteria 2). The number of positive class instances of criteria 1 is 5/12, and the number of negative class instances of criteria 2 is 7/12. Plugging this into equation 2.1 yields 0.9799. For criteria 2, the number of positive classes and negative classes is 3/4 and 1/4, respectively, yielding an entropy of 0.81128. The total entropy is then found by multiplying the number of instances A is ≥5 by the entropy found in criteria 1, plus the number of instances A is <5 by the entropy found in criteria 2.

- For feature A ≥ 5 and class == positive: $\frac{5}{12}$
- For feature A ≥ 5 and class == negative: $\frac{7}{12}$
- $Entropy(5,7) = -1[(\frac{5}{12})\log_2(\frac{5}{12}) + (\frac{7}{12})\log_2(\frac{7}{12})] = 0.9799$
- For feature A <5 and class == positive: $\frac{3}{4}$
- For feature A <5 and class == negative: $\frac{1}{4}$
- Entropy(3,1) = -1[$(\frac{3}{4})\log_2(\frac{3}{4}) + (\frac{1}{4})\log_2(\frac{1}{4})$] = 0.81128
- Entropy(Target, A) = P(≥5) × E(5,7) + P(<5) × E(3,1) = $\frac{12}{16} \times 0.9799 + \frac{4}{16} \times 0.81128$ = 0.937745

The information gain can then be found by subtracting the entropy of the target (1 because this is binary classification), by the entropy of feature A, yielding in this example 0.062255. Doing this for all four features (features), feature B is found to have the highest information gain of 0.707095, therefore by this method, it is known as the "best" feature and should be placed at the root node of the decision tree.

The gini index is used to measure how often a randomly chosen element would be incorrectly identified; therefore the feature with the lowest gini index is the best. The gini index is found for each criterion (as was done for the entropy detailed above), multiplied by the number of instances each criterion occurs within the dataset, and added to one another. Following the example above, for the same dataset with same criterion, the gini index for feature A can be found:

- For feature A ≥ 5 and class == positive: $\frac{5}{12}$
- For feature A ≥ 5 and class == negative: $\frac{7}{12}$
- gini(5,7) = 1 - $[(\frac{5}{12})^2 + (\frac{7}{12})^2]$ = 0.4860
- For feature A <5 and class == positive: $\frac{3}{4}$
- For feature A <5 and class == negative: $\frac{1}{4}$
- gini(3,1) = 1-$[(\frac{3}{4})^2 + (\frac{1}{4})^2]$ = 0.375
- gini(Target,A) = $(\frac{12}{16})\times(0.486) + (\frac{4}{16})\times(0.375)$ = 0.45825

Performing this on all four features, feature (feature) C yields the smallest gini index of 0.2, and therefore by this criteria would be chosen as the "best" feature to place at the node of the decision tree.

In summary, a decision tree can be built using two steps. The first being tree construction,

where the user splits the tree according to selected features, by first selecting the "best" feature for the root node, and secondly tree pruning, which consists of identifying and removing irrelevant branches (such as those that lead to outliers) to increase the classification accuracy. Disadvantages to the decision tree algorithm include a high probability of overfitting and a lower prediction accuracy as compared to other machine learning architectures, which is where the adaptation of this algorithm to a Random Forest becomes more useful in binary classification problems.

Random Forest

As mentioned, there are limitations to the decision tree model, which tends to have high variance and is known for overfitting, leading to poor performance on the testing data. To combat for high variance and overfitting, the concept of Random Forest (RF) was introduced. A random forest is a type of classifier made up of many decision trees. The fundamental idea behind a random forest is to utilize an ensemble method of classification. It combines many decision trees into a single model and combines predictions made by these trees, thereby yielding improved predictive results when compared to a single decision tree. In the case of random forest, the model creates a "forest" of random, uncorrelated decision trees to arrive at the "best" prediction [67]. There are two particular instances of randomness in a random forest model, first in selecting the data observations to be used in the forest, and secondly in selecting a random set of features at each node to be evaluated.

Combining the trees, a random forest trains each tree on a slightly different set of data observations, splitting the nodes in each tree on a limited number of features (chosen randomly). The final predictions of the random forest model are made by averaging the predictions of each tree. A random forest can be trained, for some number of trees, T, as follows:

1. Sample N cases at random with replacement (known as bootstrapping), which means

that some samples will be used multiple times in a single tree, to create a subset of the data (for training, should be ~70-80% of the total dataset)

2. At each node:

   (a) For some number *m*, *m* features are selected at random from all possible features

   (b) The feature that provides the best split, according to the methodology chosen (i.e. gini or information gain), is used to do a binary split on that node

   (c) At the following node, choose another *m* features and repeat

Figure 2.7 depicts an example of Random Forest architecture, with X representing the feature set representing the data of interest.

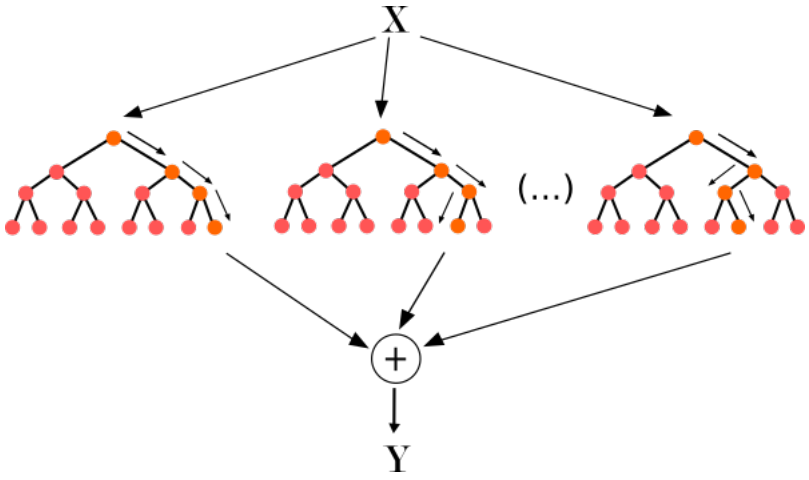

Figure 2.7: Random Forest Architecture [7]

Extremely Randomized Trees

As the name suggests, Extremely Randomized Trees (ERT) is a tree-based ensemble method, building off of the randomized forest architecture. The methodology behind this algorithm is to build totally randomized trees, whose structures are independent of the target variable values of the learning sample [8]. The two main differences of ERT as compared to other tree-based ensemble methods is that ERT:

1. splits nodes by choosing cut-points fully at random;

2. uses the whole learning sample to grow the trees (no boostrapping as in RF).

Choosing cut-points at random and using the entire learning sample to grow the trees introduces extra randomization as compared to the Random Forest model because instead of selecting an feature to split the node on based on the gini or information gain, the splitting feature is selected randomly. This methodology should be able to reduce variance more strongly than other, weaker randomization schemes. From a bias point of view, the usage of the full original learning sample as opposed to bootstrap replicas can minimize model bias. From a computational point of view, the complexity of the tree growing procedure is on the order of $N\log N$, where N is the learning sample size. Guerts, P. et. al. [8] describe the ERT algorithm well through pseudocode, shown in Figure 2.8 below:

Table 1 Extra-Trees splitting algorithm (for numerical attributes)

Split_a_node(S)
Input: the local learning subset S corresponding to the node we want to split
Output: a split $[a < a_c]$ or nothing
– If **Stop_split**(S) is TRUE then return nothing.
– Otherwise select K attributes $\{a_1, \ldots, a_K\}$ among all non constant (in S) candidate attributes;
– Draw K splits $\{s_1, \ldots, s_K\}$, where $s_i = \textbf{Pick\_a\_random\_split}(S, a_i), \forall i = 1, \ldots, K$;
– Return a split s_* such that $\text{Score}(s_*, S) = \max_{i=1,\ldots,K} \text{Score}(s_i, S)$.

Pick_a_random_split(S,a)
Inputs: a subset S and an attribute a
Output: a split
– Let $a^S_{\max}$ and $a^S_{\min}$ denote the maximal and minimal value of a in S;
– Draw a random cut-point a_c uniformly in $[a^S_{\min}, a^S_{\max}]$;
– Return the split $[a < a_c]$.

Stop_split(S)
Input: a subset S
Output: a boolean
– If $|S| < n_{\min}$, then return TRUE;
– If all attributes are constant in S, then return TRUE;
– If the output is constant in S, then return TRUE;
– Otherwise, return FALSE.

Figure 2.8: Extremely Randomized Trees Algorithm [8]

Support Vector Machine

A Support Vector Machine (SVM) classifier aims to identify elements in each class by designing a hyperplane to divide the data into classes. Considering Figure 2.9, note that there are two classes of interest in this example (a case of binary classification), red squares and blue circles, and there are two features, x_1, and x_2. The green lines in Figure 2.9 show the possible hyperplanes that could exist to separate each class accordingly.

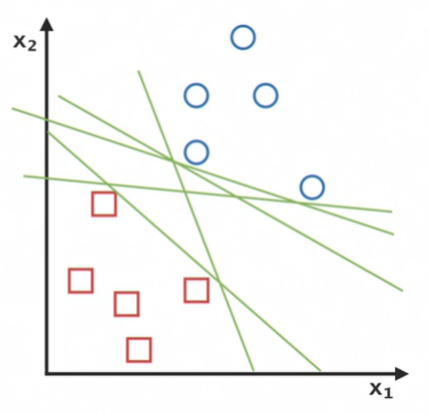

Figure 2.9: Two-dimensional Support Vector Machine Example. [9].

The goal of the SVM classifier is to optimize the hyperplane such that the margin between the decision boundary (aka the hyperplane line shown in green) and the closest points from each class is maximized.

In SVM, the data must be linearly separable, however this is not always the case. For data that is not linearly separable, the kernel method can be used. The kernel method transforms the data that is not linearly separable in an n-dimensional space, to a higher dimension where it is separable [9]. Therefore, one could project data from a 2-D space where the hyperplane (or decision boundary) is represented as a line, to a 3-D space when the hyperplane is represented as a plane (Figure 2.11). In essence, the hyperplane should be designed such that it can capture the behaviour of the data, by separating the data into two classes (in the case of binary classification) via a decision boundary.

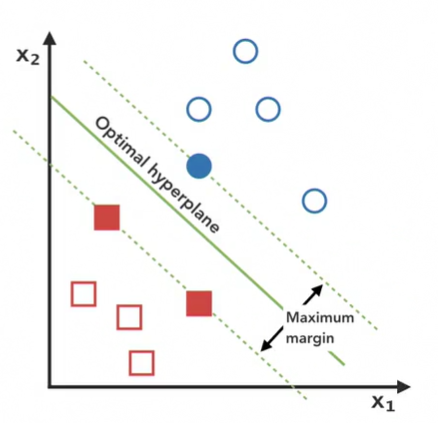

Figure 2.10: Two-dimensional Support Vector Machine Example. [9].

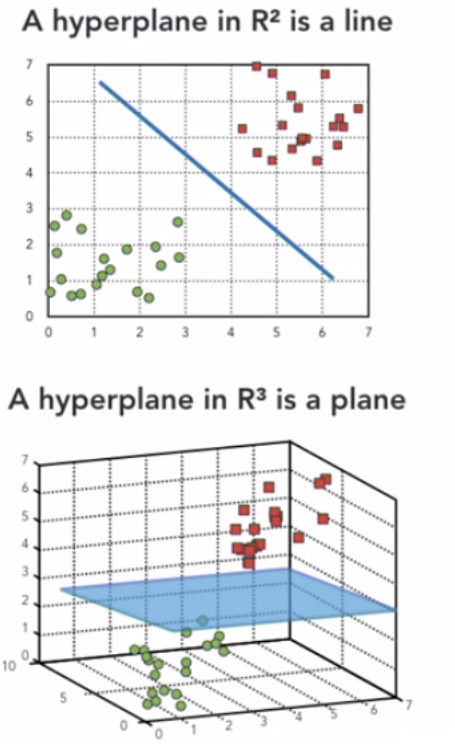

Figure 2.11: SVM Line to Plane [9].

Logistic Regression

Regression is a statistical process for estimating relationships between variables, often used to make a prediction about an outcome [9]. Logistic Regression (LR) is a multivariable method for modeling the relationship between independent variables (in the case of machine learning,

known as features) and dependent variables (classes) [68]. The model that describes the relationship between the features and class expresses the predicted value of the class variable as the sum of products, where each product is formed by multiplying the feature value and the coefficient of the feature variable. The feature variable coefficient is obtained through a mathematical fit. The relationship between the feature and class variables is represented by an S-shaped Sigmoid curve. Using the Sigmoid curve, the LR function provides estimates between 0 and 1. Logistic Regression is preferred over linear regression because of the Sigmoid curve describing the relationship between independent and dependent variables. For example, in linear regression, the line of best fit is denoted as:

$$y = mx + b \tag{2.2}$$

Where y, x represent the x and y-coordinates of each point, m is the slope of the line, and b the y-intercept. In logistic regression, the line of best fit is a Sigmoid curve represented by the equation:

$$y = \frac{1}{1 + e^{-(mx+b)}} \tag{2.3}$$

Figure 2.12 depicts the difference between linear and logistic regression, and visually shows why logistic regression is better suited for a binary classification problem (The data points at $y = 0$ represent the "0" class, and the data points at $y = 1$ represent the "1" class). The straight line (linear regression) attempts to fit all of the data to the line, which results in impossible predictions greater than one and less than zero.

The Sigmoid curve starts with a slow linear growth followed by exponential growth, ending again with a slow linear growth, and fits the data to the curve much better than the linear regression technique.

Logistic Regression modelling operates under several assumptions. The first is that LR can handle non-linear relationships between dependent and independent variables, because it ap-

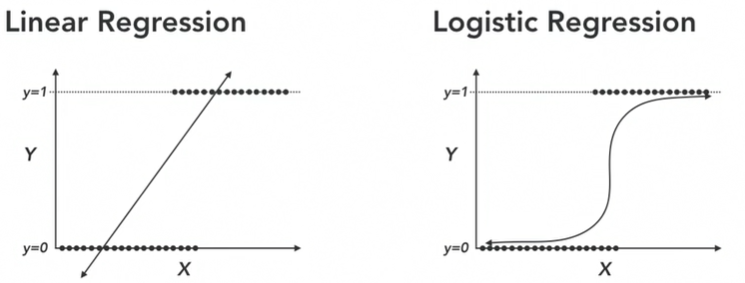

Figure 2.12: Linear vs Logistic Regression [9].

plies a non-linear logarithmic transformation to the linear regression model [68]. LR requires each observation to be independent, and independent variables should not be linear functions of one another, but rather the independent variables should be linearly related to the logarithmic probabilities of an event. Finally, LR requires large sample sizes because it uses maximum likelihood estimation.

2.5.4 Classification Performance Metrics

Confusion Matrix

A confusion matrix divides a classifier's predictions into correct and incorrect categories. A True Negative (TN) and True Positive (TP) represent a prediction that was correctly identified as a negative event and positive event respectively. A False Negative (FN) and False Positive (FP) represent a prediction that was incorrectly identified as either a negative event or positive event, respectively. A confusion matrix is a convenient way to display the classifier's performance. Figure 2.13 depicts the structure of a confusion matrix, where in practice, "True Positive", "True Negative", "False Positive", and "False Negative" would be replaced with the number of instances a classifier identified each category.

Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Figure 2.13: Confusion Matrix [10]

Accuracy

Accuracy is the most common metric to test classifier performance, and is defined as the fraction of samples correctly predicted:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

Recall

The recall of a classifier is also known as it's sensitivity, described as the fraction of positive events that were predicted correctly:

$$Recall = \frac{TP}{TP + FN} \tag{2.5}$$

Precision

The precision of a classifier is the fraction of positive events identified by the classifier that are correctly identified as positive, described by:

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

F1 score

The F1 score of a classifier is the harmonic mean of the recall and precision, described by:

$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{2.7}$$

Classifiers aim for the highest F1 score possible which represents a better performing model.

Receiver Operating Characteristics Curve

The Receiver Operating Characteristics (ROC) curve is a probability curve which tells the capability of the classification model to distinguish between classes, with the True Positive Rate on the y-axis, and the False Postive Rate on the x-axis. The Area Under the ROC (AUC) curve is a percentage value between 0 and 1. The higher the AUC, the better the model is at predicting the 0 class as 0's and the 1's class as 1's. Figure 2.14 depicts an example of an ROC curve. The dashed line represents an AUC = 0.5, an ROC curve for random guessing, and should be used as a baseline threshold for a model to perform above. With an AUC of 0.5, the model is not able to separate the classes, or has no class separation capacity, and performs no better than random guessing

Brier Score Loss

For a set of N samples, the Brier score measures the mean squared difference between the predicted probability assigned to the possible classification for each item, and the actual outcome. The smaller the Brier score, the better; the lower the Brier score is for a set of predictions, the better the predictions are calibrated [69]. The Brier score is appropriate for binary classification, and is composed of refinement loss and calibration loss. The score is always between 0 and 1, because the largest difference between a predicted probability (0-1) and the actual

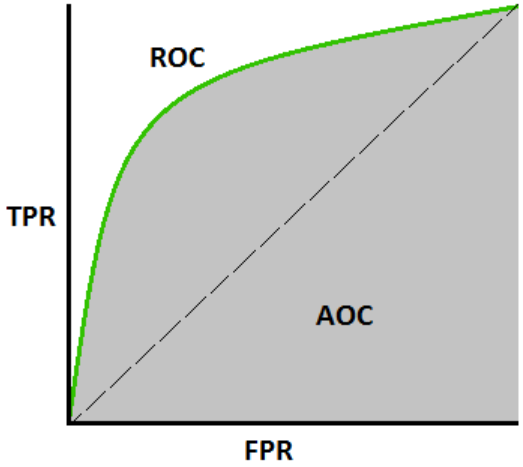

Figure 2.14: Confusion Matrix [11]

outcome (0-1) cannot be greater than 1.

Chapter 3

Classification of Accelerometer Data

At the time of this thesis completion, there has been no published literature using unprocessed satellite accelerometer data in machine learning for space weather applications. However, more and more work has been done in the fields of utilizing accelerometer and other motion sensor data to identify human activity and behaviour, and the techniques used in these works were utilized within this thesis. It has been shown by several authors (Leutheuser et. al. [70], Shoaib et. al. [71] and J.L. Reyes-Ortiz [72]) that machine learning classification techniques can be used for human activity recognition and provide accurate results.

3.1 Related Work

3.1.1 Classification of Human Activity with Accelerometer Data

Recent work done by Siirtola and Röning [72] used accelerometer data captured from users cellphones to delineate between five different human activities: idling, walking, cycling, running, and driving. The authors used two types of classifiers, k-nearest-neighbours (knn) and quadratic discriminant analysis (QDA). They used a tri-axis accelerometer with a sampling

frequency of 40Hz. A total of 21 features were extracted from the magnitude acceleration sequences, standard deviation, mean, minimum, maximum, five different percentiles (10, 25, 50, 75, and 90), and a sum and square sum of observations above/below certain percentile (5, 10, 25, 75, 90, and 95). The same features were extracted from the signal where 2/3 acceleration channels were square summed together, generating a total of 42 features; 21 from the magnitude acceleration signal and 21 from the signal where the x and z component were square summed. The average recognition rate using QDA was 0.958, and the knn performed slightly worse with an average recognition rate of 0.939.

Most notably, the work done by E. Zdravevski et al. [12] heavily influenced the approach undertaken in this research, and will be detailed in the following subsections.

3.1.2 Benchmark Classification Model

E. Zdravevski et al. [12] looked at an approach to *automatic machine-learning based identification of jogging periods from accelerometer measurements*, using a dataset containing accelerometer measurements placed on the hip and ankle of 39 fifteen year old adolescents. The authors used four different types of classification algorithms, Random Forest, Extremely Randomized Trees, Support Vector Machines, and Logistic Regression. Zdravevski et al.'s work will be used as the benchmark of classifier performance for this thesis. Figure 3.1 depicts a sample of the accelerometer readings used in the study. Note that any errors in the labels or incorrect alignment with the recorded data (such as unsynchronized time) were fixed by the author and relabelled as golden, whereas the original entries from the participants have diary a part of their time series labels, respectively.

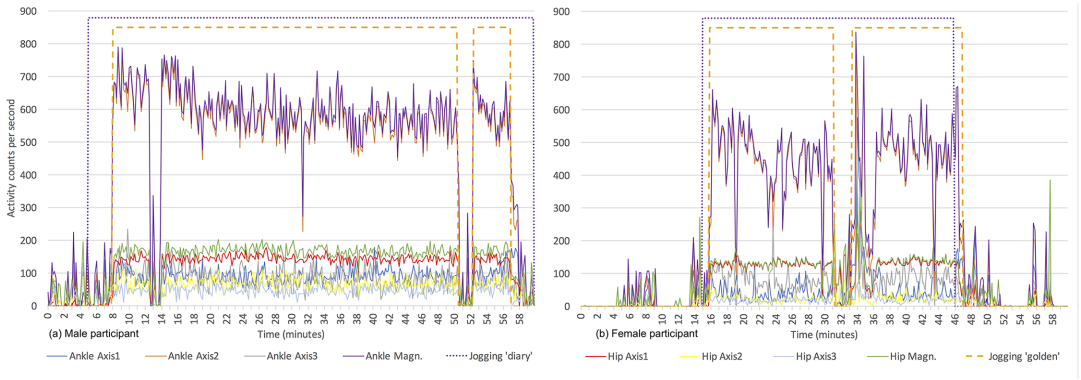

Figure 3.1: Raw accelerometer readings for one hour during which two participants (a) male and (b) female, had a jogging activity [12]

3.1.3 Feature Selection

Feature selection was done manually, divided into six categories: basic statistics, equal-width histogram, percentile-based, correlations, auto-correlations, and curve fitting parameters. Features were extracted from twelve original time series with data segmented based on two strategies: 60 second windows without overlap, and 180 second with 120 second of overlap, in addition to time series derived from the original from both hip and ankle accelerometers. A flow diagram of the feature extraction, selection, and classification process is shown in 3.2.

3.1.4 Classifier Performance

Figures 3.3 and 3.4 show the performance of the four different classifiers with the four final feature sets for both the 60 second non-overlapping windows and 180 second overlapping windows. In both scenarios, all of the classifiers have a very high accuracy (over 0.995). It is important to note that it is not expected that the models used in this study will reach the level of performance of Zdravevski et al.'s, due to their much greater data volume, multi-axis accelerometer data availability, and supported literature for human activity recognition using accelerometer data. However, the fundamental idea and strategy of the two works is the same, and their results show well-performing classification ability.

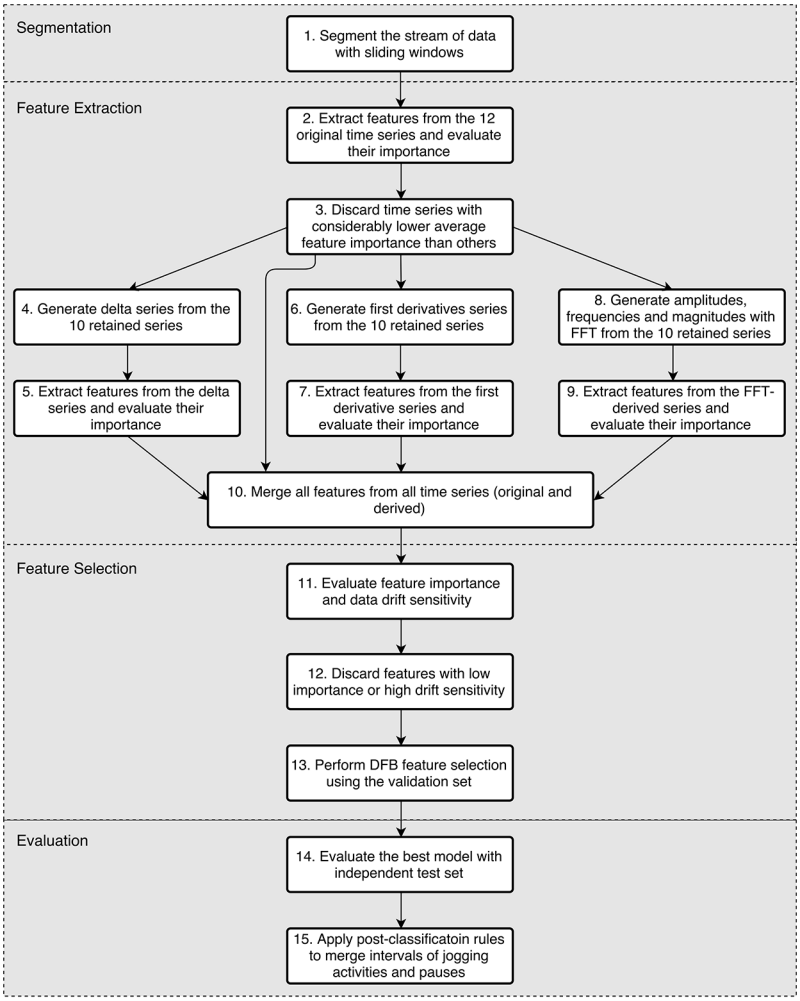

Figure 3.2: Algorithm for feature extraction, selection and classification [12]

Features	Classifier	Acc.	AUC	Prec.	Recall	Spec.	F1	Time
Best Ankle (8 feat.)	ERT	0.9970	0.9906	0.9339	0.8937	0.9988	0.9133	4.0
	RF	0.9969	0.9883	0.9579	0.8633	0.9993	0.9081	4.7
	Logistic	0.9967	0.9986	0.9259	0.8861	0.9987	0.9056	0.1
	SVM	0.9917	0.9940	0.7137	0.8962	0.9934	0.7946	24.3
Best Hip (20 feat.)	ERT	0.9967	0.9951	0.9171	0.8962	0.9985	0.9065	5.5
	RF	0.9962	0.9906	0.8939	0.8962	0.9981	0.8951	6.9
	Logistic	0.9977	0.9991	0.9528	0.9190	0.9992	0.9356	0.4
	SVM	0.9970	0.9988	0.9533	0.8785	0.9992	0.9144	188.2
All (17 feat.)	ERT	0.9954	0.9977	0.8600	0.8861	0.9974	0.8728	9.9
	RF	0.9959	0.9959	0.8961	0.8734	0.9981	0.8846	8.2
	Logistic	0.9963	0.9911	0.9023	0.8886	0.9982	0.8954	0.9
	SVM	0.9954	0.9987	0.8391	0.9241	0.9968	0.8795	249.9
Best Ankle + Best Hip(28 feat.)	ERT	0.9971	0.9966	0.9321	0.9038	0.9988	0.9177	7.8
	RF	0.9972	0.9911	0.9514	0.8911	0.9992	0.9203	7.7
	Logistic	0.9968	0.9992	0.9030	0.9190	0.9982	0.9109	0.5
	SVM	0.9976	0.9993	0.9525	0.9139	0.9992	0.9328	204.7

Figure 3.3: Classifier results 60 second windows [12]

Features	Classifier	Acc.	AUC	Prec.	Recall	Spec.	F1	Time
Best Ankle (17 feat.)	ERT	0.9990	0.9930	0.9571	0.9750	0.9993	0.9659	5.0
	RF	0.9988	0.9928	0.9509	0.9688	0.9993	0.9598	6.1
	Logistic	0.9989	0.9994	0.9837	0.9406	0.9998	0.9617	0.3
	SVM	0.9994	0.9972	1.0000	0.9625	1.0000	0.9809	25.2
Best Hip (20 feat.)	ERT	0.9981	0.9976	0.9761	0.8938	0.9997	0.9331	7.1
	RF	0.9978	0.9975	0.9564	0.8906	0.9994	0.9223	5.6
	Logistic	0.9967	0.9988	0.8739	0.9094	0.9980	0.8913	0.4
	SVM	0.9968	0.9994	0.8862	0.9000	0.9983	0.8930	141.4
All(12 feat.)	ERT	0.9954	0.9977	0.8600	0.8861	0.9974	0.8728	9.9
	RF	0.9975	0.9935	0.9102	0.9188	0.9987	0.9145	5.5
	Logistic	0.9990	0.9962	1.0000	0.9344	1.0000	0.9661	0.3
	SVM	0.9972	0.9964	0.8862	0.9250	0.9982	0.9052	24.8
Best Ankle + Best Hip(37 feat.)	ERT	0.9988	0.9982	0.9399	0.9781	0.9991	0.9587	7.6
	RF	0.9983	0.9964	0.9354	0.9500	0.9990	0.9426	7.6
	Logistic	0.9985	0.9997	0.9470	0.9500	0.9992	0.9485	1.6
	SVM	0.9984	0.9998	0.9613	0.9313	0.9994	0.9460	186.9

Figure 3.4: Classifier results 180 second windows [12]

Chapter 4

Methodology

The methodology of this thesis work from data acquisition to the final classifier model architecture can be described in the following overview, and will be detailed thoroughly in this chapter:

1. Data acquisition of the NASA JPL Level1B accelerometer data of the GRACE-A satellite;
2. First segmentation of daily data from 2002-2017 to dates with ICME storms as per the Richardson-Cane ICME list;
3. Pre-processing the raw data files to containing only the x-component of linear acceleration, time, and date (linear interpolation performed where necessary);
4. Class balancing to achieve a total dataset containing 50% ICME storms and 50% quiet days, 24-hour periods;
5. Feature extraction;
    (a) Basic statistical features with percentiles extraction;
6. Random Forest methodology testing on 24-hour ICME periods, where at least 16/24

hours contain an ICME storm (regardless of Dst strength and Kp at the time of the storm);

7. Second segmentation of data, selecting 8-hour windows of ICME storms where at least 4/8 hours have a Dst index value less than -50, and 8-hour windows with Dst greater than -50 and Kp of 4 or less for the quiet periods;
8. RF, ERT, SVM, LR testing using basic statistical features with percentiles only with 10-fold cross validation (and testing on unseen data);
9. First-order derivative of data and extraction of basic statistical features;
10. Random Forest, ERT, SVM, LR testing using basic statistical features and first-order derivative features with 10-fold cross validation (and testing on unseen data);
11. Choosing RF and ERT architectures for further performance improvement;
12. Feature importance ranking and rejection of unimportant features of RF and ERT;
13. Hyperparameter tuning of RF and ERT using important features;
14. Selecting best hyperparameter settings to obtain final RF and ERT model and;
15. Testing RF and ERT final model on unseen test data (10% of data).

4.1 Data Acquisition

4.1.1 NASA JPL Level-1B Data

The data used in this study is the Level-1B product data available from NASA Jet Propulsion Lab, with detailed descriptions found in [2]. The Level-1B data is derived from processing applied to both the Level-1A and Level-0 data. The Level-0 data products are the result of the telemetry data reception, collection, and decommutation by the GRACE Raw Data Center at DLR in Neustrelitz. The Level-1A data is the result of non-destructive processing applied to the

Level-0 data; sensor calibration factors are applied to convert the binary encoded measurements to engineering units. Time tag integer second ambiguity is resolved and data is time tagged with respect to the satellite receiver clock time, and editing and quality control flags are added. The Level-1A data can be reversible to Level-0, barring any bad data. The Level-1B data is then obtained through further processing, correctly time-tagging the data, and reducing the sample rate. Level-1B includes the ancillary data products generated during processing, and additional data needed for further processing. More information on the exact algorithms used for processing can be found in Sien-Chong Wu & Gerhard L.H. Kruizinga [73].

The accelerometer data used in this study is known as ACC1B, which provides the linear and angular acceleration components of the proof mass of the GRACE satellites. Due to the orientation of the spacecraft along its orbit path, only the x-component of linear acceleration is used in this study, as this is the acceleration component of the satellite that captures change due to space weather events. The SuperSTAR accelerometer used by GRACE-A has a highly-sensitive along-track (x) axis, which is the direction approximately tangential to its orbit path [18]. In the along-track direction, the main contribution to the non-conservation acceleration is caused by atmospheric drag, it therefore makes sense that the SuperSTAR accelerometer would capture changes in acceleration caused by an ICME event, since ICMEs cause perturbations to the atmosphere's neutral density, thereby altering the drag force on GRACE-A. The drag acceleration is described by:

$$\vec{a} = -\frac{1}{2}\rho v_r^2 C_D \frac{A}{M}\frac{\vec{v_r}}{||\vec{v_r}||} \tag{4.1}$$

where ρ is the atmospheric density, $\vec{v_r}$ the orbital velocity of the spacecraft, C_D the drag coefficient, A the equivalent cross-sectional surface perpendicular to the direction of motion and M the spacecraft mass. Any increase of the local density (ρ) of the atmosphere induces an increase of the drag force on a spacecraft, captured by the SuperSTAR accelerometer onboard GRACE. An example of the accelerometer data during an ICME vs quiet period is shown in

Figures 4.1 and 4.2.

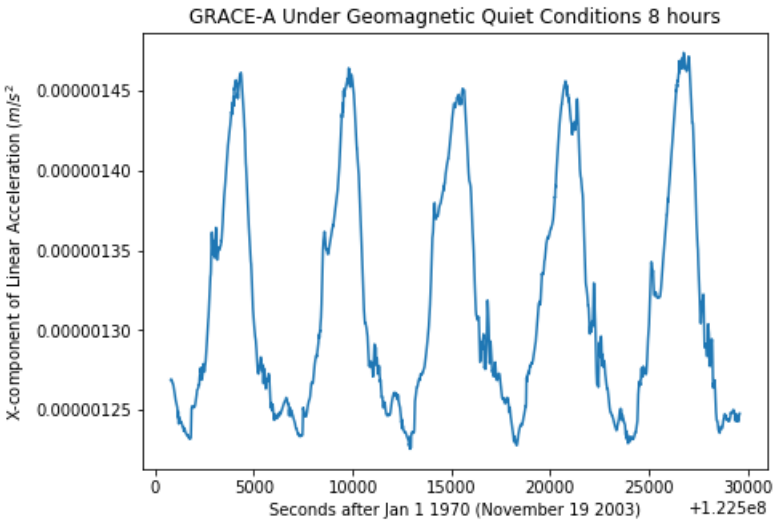

Figure 4.1: GRACE-A Acceleration during 8-hour Quiet Period November 2003

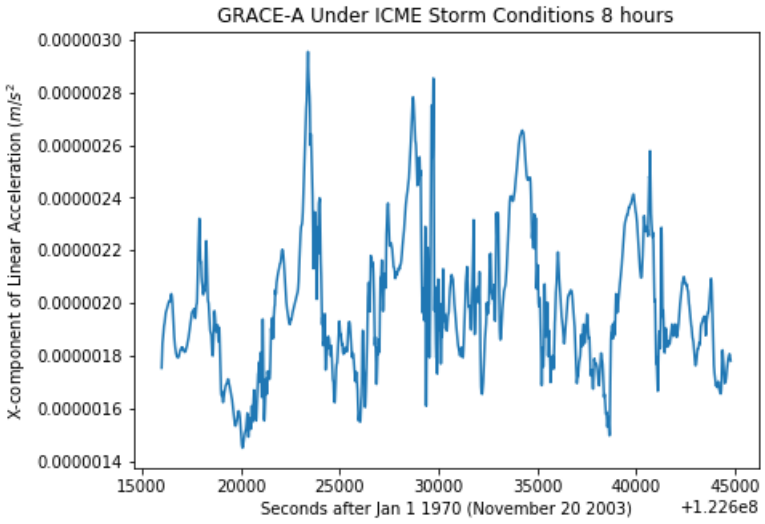

Figure 4.2: GRACE-A Acceleration during ICME Storm 8-hour Period November 2003

Visually, there is a clear difference between the accelerometer behaviour during an ICME storm, and during a period of quiet geomagnetic activity.

Using the Richardson-Cane ICME list from 1996-2019, the ICME storms were cross-listed with the available accelerometer data over the lifetime of the GRACE-A satellite. The database provides estimated start and end times of the ICME storm based on plasma observations. In

order to choose "quiet periods" that would account for seasonal changes of density in the atmosphere, every Wednesday from 2002-2017 (if ICME present on this day, then the next closest day was chosen) was selected as a reference quiet period (aka, a non-event as would be interpreted by the classifier). Therefore, 4 days per month, for every month from 2002-2017 over the lifetime of the GRACE-A mission (data availability permitting) was labelled as a "0" event. To confirm that there was no other geomagnetic activity on that day (in case it was missed by the Richardson-Cane list) the Dst index of that day was checked via [74]. Dst index is a good indication for geomagnetic activity caused by ICME storms. In addition, the Kp was also checked. "Quiet days" were rejected if their Dst index at any time during that day had a magnitude larger than 50, and their Kp index greater than 4+, if this occurred, the next closest day that was not during an ICME storm was chosen.

4.2 Class Balancing

The available GRACE-A accelerometer data is from April 4th, 2002 until June 29th, 2017. Within those dates, there are some days missing, likely due to spacecraft maintenance that was required once the satellite was over its expected mission lifetime of 5 years. Considering these dates, and rejecting those where data is unavailable, the ICME storms as listed by Richardson and Cane that are captured by the dataset total 291. The maximum duration of an ICME is 90 hours, and the minimum duration is 3, with an average storm length of ~24 hours. In total, there are ~7161 hours of ICME storms within this dataset. Immediately, this presents a class imbalance problem; if all the ICME storms are taken to be 24 hours in length, or one full day, this means there are ~291 days within the dataset of ICME storms, or the "1" class. The entire dataset contains ~5066 days, which means that there are 291 positive class instances, and 4775 negative class instances (days without an ICME storm), resulting in the ICME storms making up only ~6.1% of the data. To combat this class imbalance, a down-sampling technique was used.

4.2.1 Down-sampling and Window Selection Iteration One

The first iteration of down-sampling and class-balancing the data involved analyzing the storms and quiet periods over 24-hour windows. Each accelerometer file encompasses a 24-hour period, with one measurement taken every second with the exception of missing data that required interpolation. GRACE had an orbital period of ~94 minutes, therefore a 24-hour period would contain approximately ~15 orbits of the satellite. Although the average storm length is ~24 hours, to capture more of the storm data for the "1" class, storms between 16-24 hours were used. Storms shorter than 24 hours were "padded" with the data immediately before and/or after the occurrence of the event. It was initially hypothesised that two thirds of a 24-hour period and greater should be sufficient enough to extract a feature vector that would maintain the integrity of the satellite's reaction to a storm event. In total, this method encompassed 69 ICME storms between 16-24 hours. For a 50/50 class balance, one quiet period for every ICME around the same time of year was taken. The newly-segmented, balanced dataset was first used in a Random Forest architecture to test its feasability prior to testing on all four classifier architectures, with details described in the Results section.

4.2.2 Down-sampling and Window Selection Iteration Two

Following the poor performance of the Random Forest 24-hour ICME window classifier, the accelerometer data was visually inspected, to fully understand the reasoning behind the performance. Taking the previous examples during November 2003 shown in Figures 4.1 and 4.2, there is a clear visual difference between the 8-hour periods. However, upon inspection of the entire 24-hour period the visual is much different, shown in Figures 4.3 and 4.4:

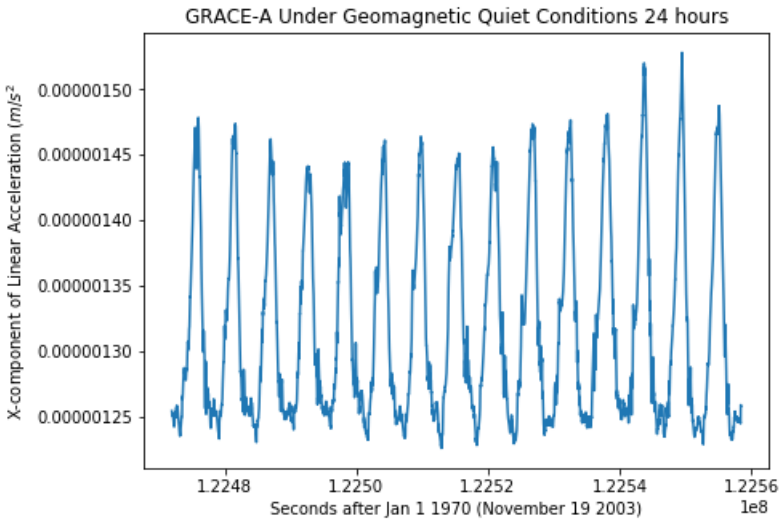

Figure 4.3: GRACE-A Acceleration during 24-hour Quiet Period November 19th 2003

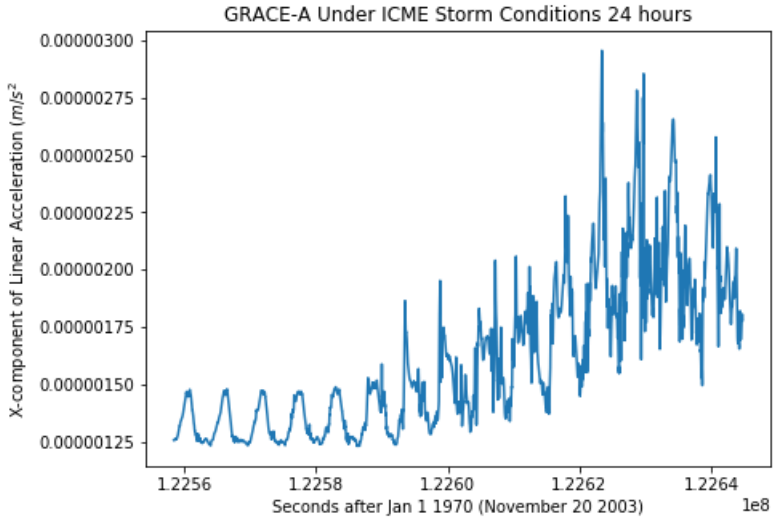

Figure 4.4: GRACE-A Acceleration during ICME Storm 24-hour period November 20th 2003

Approximately one third of Figure 4.4 exhibits a period of quiet geomagnetic activity, which makes sense as the ICME storm for this particular date lasts from 10:00:00 on 2003/11/20 - 8:00:00 2003/11/21 (UTC), therefore the quiet acceleration pattern exhibited at the beginning of the figure corresponds with the satellite's behaviour prior to it's perturbed state during the storm. For the above example shown, statistical features such as the mean, range, and maximum should be different enough between the two cases (quiet vs storm) that it can be easily differentiated and therefore predicted by a classifier. However, consider the following example

of September 19th-21st, 2005 shown in Figures 4.5 and 4.6.

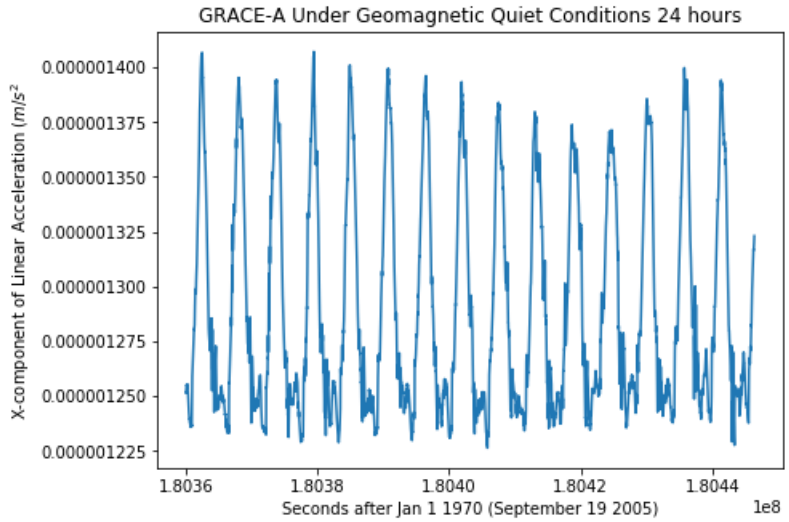

Figure 4.5: GRACE-A Acceleration during 24-hour Quiet Period September 19th 2005

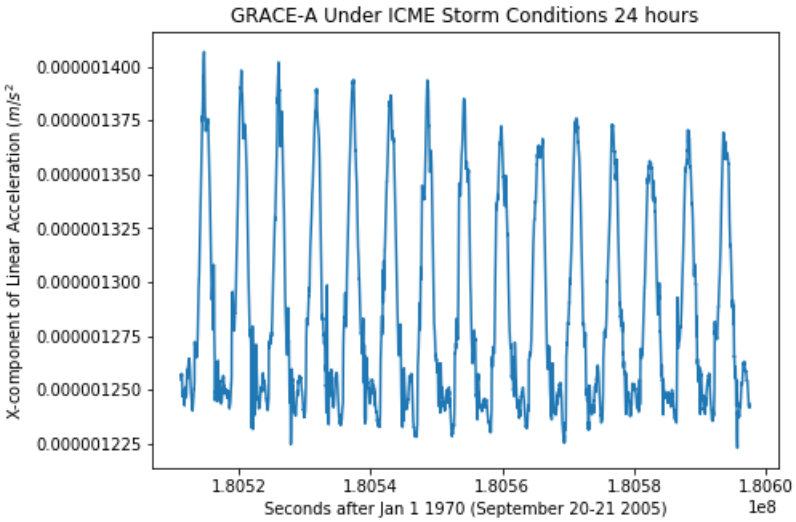

Figure 4.6: GRACE-A Acceleration during ICME Storm 24-hour period September 20th-21st 2005

From 18:00:00 2005/09/20- 18:00:00 2005/09/21 there was a recorded ICME storm according to Richardson Cane. The Dst index during this period, however, did not fall below -33; hardly low enough to classify this period as under the influence of a geomagnetic storm. Furthermore, the Kp index for these dates does not exceed 2, which denotes a period of quiet activity. Although there is an ICME storm lasting 24 hours during September 20-21 2005,

the geo-effectiveness, or strength of the storm, is negligible from an atmospheric and satellite perturbation point of view. This is supported by the visual evidence of Figures 4.5 and 4.6 which show the GRACE-A satellite behaving similarly during a period of quiet activity and during a storm event. The September 2005 example, among many others over the lifetime of GRACE-A, was the motivation for the second down-sampling methodology described below.

Though quiet periods in iteration one of down-sampling had Dst and Kp thresholds (i.e., the 24-hour period in question must have its Dst and Kp fall between certain values to be a valid "quiet" period), the same was not done for the ICME periods. This resulted in the poor performance of the RF model as storms that did not cause accelerometer perturbations were included in the dataset. To mitigate this, and therefore only include those storms that were likely "strong" enough to cause accelerometer perturbations, Dst and Kp thresholds for ICME periods were imposed. The Dst index throughout the duration of the storms was checked via the World Data Center for Geomagnetism, Kyoto [74] to improve likelihood of classifier success, and to include only the ICME storms as defined by Richardson and Cane that are likely strong enough to cause significant geomagnetic disturbance detectable by GRACE. To encapsulate the maximum storm strength of the available ICMEs, 8-hour periods (or windows) of the strongest consecutive values of Dst index were chosen for each storm. Storms were rejected if at least 4 of the 8 hours of the highest Dst index were above -50. From this, 87 ICME storms with available accelerometer data met the criteria and were kept in the dataset for training and testing. Next, down-sampling of the "quiet" data was done. To meet a desired 50/50 class balance, the same number of "quiet days" as ICME storms were chosen. The criteria for a "quiet day", or the "0" class, was the following: days were chosen based on their closest proximity to an ICME storm (2-4 days if possible), with 8-hours of data with a Dst index greater than -50 for the entire duration of the period, and a Kp index of less than 5. The maximum Kp value for the quiet periods used in this study was 4+, which is characterized as "Active", and may have bright, constant and dynamic northern lights visible in the atmosphere at this time [75].

Following the class-balancing, a 50/50 split of ICME storms and quiet days leave 174, 8-hour period samples to train and test a classifier with, 87 files per class. The limitations to the machine learning model and space weather prediction capability using the described criteria is discussed in Chapter 6.

4.3 Pre-processing

4.3.1 Working with the raw files

The raw .dat files from JPL were first converted to ascii and the headers were removed, then converted to .txt files. The time data within each file had to be converted from the JPL epoch of "seconds after January 1, 2000, 12:00:00PM UTC" to the generally used epoch of January 1, 1970, 00:00:00 UTC, for use in the Python language and avoidance of complication and confusion when manipulating datetime variables. Daylight savings time was also taken into consideration to ensure correct conversion. Following the conversion and file manipulation to include only information of interest to this study, the accelerometer files contained the x-component linear acceleration of GRACE (as the remaining two do not show variation caused by solar activity due to their relative position on the satellite orbit path), seconds past January 1, 2000 12:00:00 PM epoch, seconds past January 1, 1970 00:00:00 epoch, and the corresponding full date and 24 hour time. In theory, one accelerometer file should represent 86400 points of data, starting at midnight (00:00:00) until the end of day (23:59:59), with one measurement, or data point, every second. Most files are formatted this way, however there are cases where up to 300 consecutive points could be missing. When considering that 300 seconds out of a total of 86400 are missing, this is 0.3% of the time series. Interpolation to fill the gaps was deemed to be an acceptable method in order for each file to be an equal-length time series because this is a small fraction of the entire day of data. JPL described in their Level 1B Handbook that "Accelerometer level 1A (ACC1A) data are pre-processed and quality flags are checked. Data

gaps of 10 seconds or shorter are filled using quadratic interpolation when at least 2 points per side are available. If there are less than two points per side, then linear interpolation is used to fill the data gap. Data gaps exceeding 10 seconds are not filled" [2]. Therefore, linear interpolation over the gaps was deemed to be sufficient in keeping the integrity of the dataset intact.

4.3.2 Extracting Feature Vectors

When extracting the feature vectors from the accelerometer data, [12]'s work was followed extensively. Their results show a classification accuracy of at least 0.99, using logistic regression, support vector machines, random forest, and extremely randomized trees machine learning algorithms. Data was first normalized using a max-min normalization method, and was calculated as follows: for every value V, there is a normalised value, V_N that can be found through:

$$V_N = \frac{V - minimum}{maximum - minimum} \tag{4.2}$$

Basic Statistics with Percentiles

For Experiment one, a feature vector containing the 13 features was used. These features are based on basic statistics and percentiles, and include: minimum, standard deviation, mean, 25th percentile, 50th percentile, 75th percentile, maximum, range, median, geometric mean, variance, sum of series, and signal to noise ratio of the accelerometer data. The standard deviation of the data is described by:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4.3}$$

with N being the total sample size, x_i representing the value of one sample, μ the mean of the dataset, and σ the standard deviation.

The mean and geometric mean are calculated differently, with the mean shown in Equation 4.4 and the geometric mean in Equation 4.5:

$$\mu = \frac{\sum X_i}{N} \tag{4.4}$$

where μ represents the mean, $\sum X_i$ is the summation of the samples, and N the number of samples.

$$(\Pi_{i=1}^{N} x_i)^{\frac{1}{N}} \tag{4.5}$$

The geometric mean is found by taking the Nth root of the multiplication of all samples (x_i) in a population, where N refers to the total number of samples.

The variance of a population is defined as the sum of the squared distances of each sample in the population from the mean μ, divided by the total number of samples N:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \tag{4.6}$$

The sum of series of a population is the sum of the time series values:

$$\sum_{i=0}^{N-1} x_i \tag{4.7}$$

And the signal to noise ratio is defined as the mean of the samples divided by the standard deviation:

$$SNR = \frac{\bar{x}}{\sigma} \tag{4.8}$$

First-Order Derivative

For Experiment two, the first-order derivative of the accelerometer data was taken over each 8-hour period. For a time series with K readings collected at time t(i), 0≤ i≤ K, one can calculate K-1 first-order derivatives. The first derivative of a time series a and at time t(i), 0 less than i≤ N can be estimated as:

$$derivative(i) = \frac{reading(i) - reading(i-1)}{t(i) - t(i-1)} \tag{4.9}$$

The first derivative can also be defined in this case as taking the difference over the time series because the time series are of equal length, equally spaced data points.

From the creation of the first-order derivative time series, the same features from experiment one (basic statistics with percentiles) were extracted, with the exception of the derivative's geometric mean (as this included NaN values). Adding the first-order derivative time series resulted in a total of 25 features.

4.4 Test-Train-Validation Split

The data was split into three sets: training, testing, and validation. In Random Forest classification, each tree built uses a random sample (with replacement) of the training dataset and therefore does not necessarily need a validation set. However, because this thesis explores the use of four different machine learning architectures and their relative performance to one another, a validation set is necessary for fair comparison. The training set contains 80% of the data, and the testing and validation sets each contain 10% of the data respectively. Each classifier is first built and iterated to achieve the highest possible accuracy using K-folds cross validation with K=10, a commonly used number of folds. Once the models could not be improved further with the exception of adding more training data which was not possible, the

classifiers then used the unseen validation dataset to make predictions. The validation data was split randomly from the entire available dataset prior to training the classifiers.

4.4.1 K-folds Cross Validation

K-folds cross validation is a technique used in machine learning to evaluate the performance of a model on a sample of data. In the case of this research work, there is a limited amount of data available for training and testing a machine learning model. For small datasets, using a train-test-split method (for example, training on 80% of the data, testing on 20% of the data) can result in a large amount of variance. To combat the variance, the method of K-folds cross validation can be used. The idea of K-folds cross validation is to repeat the training and testing process K times, and average the accuracy of the model, to get a better picture of it's performance.

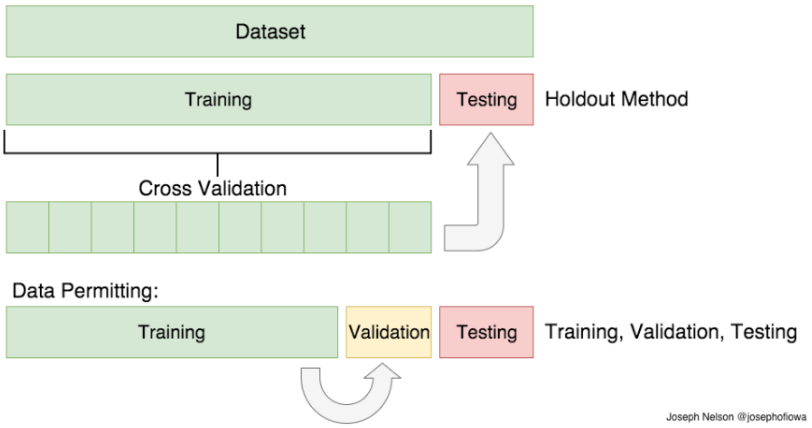

Figure 4.7: K-folds Cross Validation [13]

K-folds cross validation works as follows:

- Divide data into K-subsets (For example, let K=10);
- Use K-1 subsets (folds) to train the data, and leave the last subset as the testing data;
- For K=10, the validation set will be the first 10% of the data, use this to calculate the accuracy (and other performance metrics of interest);

- On the second iteration, the next 10% of the data will be used as the validation set, and so on. Repeat K (10) times, every time the validation set will be a different portion (subset) of the data;
- Average all the accuracy values to get the average accuracy of the model.

Figure 4.7 shows a visual representation of K-folds cross validation. The advantage of K-folds cross validation is that it matters less how the data gets divided, because every data point gets to be in a test set exactly once, and gets to be in a training set K-1 times. As K is increased, the variance of the resulting estimate is reduced [76]. The disadvantage of the K-folds cross validation method is that the training algorithm has to be rerun K times, which means it takes K times as much computation to make an evaluation. However, the dataset in question for this thesis work is small, and thus the computation time is less of a concern at this point.

The cross validation step occurs concurrently with hyperparamter tuning, such that the "best" model is found using specific metrics (i.e. number of trees, depth of each tree, etc.). Once the average performance metrics have been computed, the test set is introduced, which is a portion of the data that was set aside ahead of the cross validation step, to examine how the model performs on unseen data.

4.5 Experiment Zero: 24-hour Storm Data Hypothesis Testing

As described in the Class Balancing section, for the first iteration of segmenting the time series, 24-hour periods were used, as each accelerometer file contains one point per second for a 24-hour period (where exceptions, interpolation was used as mentioned). Only the Random Forest model was used in Experiment 0, which acted as a hypothesis-testing step. The selected features are shown in Table 4.1 below. The architecture for the random forest classifier for

experiment zero utilized mainly default values of the scikit-learn toolbox. The number of trees (estimators) was manually set to 1000. The main parameters of the random forest model are found in table 4.2.

Feature
Minimum
Standard Deviation
25th percentile
50th percentile
75th percentile
Maximum
Range
Median
Geometric Mean
Variance
Signal to Noise Ratio
Sum of Series
Mean

Table 4.1: Experiment 0 Feature Selection

Parameter	Setting
Number of trees	1000
Split criterion	gini
Maximum Depth	None
Minimum Samples Split	2
Minimum Samples Leaf	1

Table 4.2: Experiment 0 Random Forest Model Parameters

4.6 Experiment One: Basic Statistical Features

For experiment one onward, the 8-hour time series segmentation for the ICME and Quiet day data was used.

4.6.1 Selected Features

The selected features for all four models are listed in Table 4.3.

Feature
Minimum
Standard Deviation
Mean
25th Percentile
50th Percentile
75th Percentile
Maximum
Range
Median
Geometric Mean
Variance
Signal to Noise Ratio
Sum of Series

Table 4.3: Experiment 1 Features

4.6.2 Model Parameters

The model parameters used in both experiment one and two are the default settings according to the sci-kit learn toolbox. The parameter settings were kept the same for the two experiments to investigate only the change in classifier performance caused by adding more features.

Random Forest and Extremely Randomized Trees

As the Random Forest and Extremely Randomized Trees models are both decision tree-based models, their default parameter settings are the same. As a starting point, 1000 trees were used in both models. The same model parameters are used in experiment zero, one, and two for the RF and ERT models, shown in Table 4.2. Note that there is no "split criterion" for ERT, as this is done randomly.

Support Vector Machines

The scikit-learn toolbox default parameters dictated the settings used by the Support Vector Machine classifier. The SVM classifier was specified to have a linear kernel. The penalty parameter, C, was left at the default value of 1.

Logistic Regression

The scikit-learn toolbox default parameters dictated the settings used by the Logistic Regression classifier. Of particular importance, the penalty function was set to L2. L2 refers to the regularization term used by the model in order to prevent overfitting, and is the sum of the square of the model weights.

4.7 Experiment Two: Basic Statistical Features with First-Order Derivative

4.7.1 Selected Features

The selected features for all four models are the same. Experiment two consists of two time series to extract features from, the original 8-hour window of accelerometer data, and the first-order time series derivative of the 8-hour window. The same statistical features as experiment one are extracted from both time series, with the exception of geometric-mean from the first-order derivative series. In total, there are 25 features used by the models in experiment two, listed in Table 4.4 below:

Feature
Minimum
Standard Deviation
Mean
25th Percentile
50th Percentile
75th Percentile
Maximum
Range
Median
Geometric Mean
Variance
Signal to Noise Ratio
Sum of Series
Derivative Minimum
Derivative Standard Deviation
Derivative Mean
Derivative 25th Percentile
Derivative 50th Percentile
Derivative 75th Percentile
Derivative Maximum
Derivative Range
Derivative Median
Derivative Variance
Derivative Signal to Noise Ratio
Derivative Sum of Series

Table 4.4: Experiment 2 Features

4.8 Experiment Three: Feature Engineering and Hyperparameter Tuning

In experiment three, the same statistical features from the original and first-derivative time series are used. The objective of experiment three is to determine what features to keep or reject based on model performance, and to tune the hyperparameters of the Random Forest and Extremely Randomized Trees models using the best features for optimal classification accuracy, loss, and ROC-AUC curve. The results from the iterative sub-experiments within experiment three are included in Appendix C, and the final Random Forest and ERT model

performances are included in Chapter 5.

4.8.1 Random Forest

Feature Importance

Utilizing the scikit-learn feature importance ranking method, the ranked feature importance as determined by the gini impurity by the Random Forest model for the parameters was found. Table is shown below (for number of trees = 1000, 80% train, 10% test, 10% validation, cross validation = 10):

Feature	Importance
Derivative Standard Deviation	0.107413
Derivative Range	0.094405
Derivative 75th Percentile	0.083616
Derivative Minimum	0.082771
Derivative Maximum	0.077533
Derivative 25th Percentile	0.063669
Range	0.053963
Minimum	0.036193
Standard Deviation	0.035677
Variance	0.033515
Signal to Noise Ratio	0.031982
Derivative sum of series	0.029467
Derivative Signal to Noise Ratio	0.029114
Median	0.025777
Maximum	0.025644
50th Percentile	0.024477
25th Percentile	0.024371
sum of series	0.024094
75th Percentile	0.023927
Mean	0.023323
Geometric Mean	0.022465
Derivative Median	0.019208
Derivative 50th Percentile	0.017371
Derivative Mean	0.009937
Derivative Variance	0.000088

Table 4.5: Experiment 3 Random Forest Feature Ranking

Iteration 1 involved selecting features which importance is greater than the mean importance of all the features. Using the mean criteria, 7 features were retained: range, derivative minimum, derivative standard deviation, derivative 25th percentile, derivative 75th percentile, derivative maximum, and derivative range. The performance decreased from using all features, so a second iteration was explored.

Iteration 2 selected the features which importance is greater than a chosen threshold of 0.03. The threshold retained 12 features: minimum, standard deviation, range, variance, signal to noise ratio, derivative minimum, derivative standard deviation, derivative 25th percentile, derivative 75th percentile, derivative maximum, derivative range, and derivative signal to noise ratio. The performance of the model improved slightly from iteration one, but was still lower than using all features, and so a third iteration was explored.

Iteration 3 chose a feature importance threshold of 0.02, retaining 21 features: minimum, standard deviation, mean, 25th percentile, 50th percentile, 75th percentile, maximum, range, median, geometric mean, variance, signal to noise ratio, sum of series, derivative minimum, derivative standard deviation, derivative 25th percentile, derivative 75th percentile, derivative maximum, derivative range, derivative signal to noise ratio, and derivative sum of series. The performance of the model increased using this criteria, increasing accuracy by $\sim$1%, decreasing loss by 0.03%, and increasing ROC-AUC by $\sim$1%. The standard deviation of the accuracy increased by $\sim$1%, however the overall increase of the performance metrics of the model utilizing these features was deemed of greater value.

Iteration 4 chose a feature importance threshold of 0.025, retaining 16 features: minimum, standard deviation, range, median, variance, signal to noise ratio, sum of series, derivative minimum, derivative standard deviation, derivative 25th percentile, derivative 75th percentile, derivative maximum, derivative range, derivative signal to noise ratio, and derivative sum of series. The model performance decreased from iteration three, and therefore using a feature

importance threshold of 0.02, retaining 21 features, was used in the hyperprarmeter tuning step of formulating the final Random Forest classifier model.

4.8.2 Extremely Randomized Trees

Feature Importance

As with the RF model, the ranked feature importance as determined by the gini impurity by the ERT model for the parameters is shown below (for number of trees = 1000, 80% train, 10% test, 10% validation, cross validation = 10):

Feature	Importance
Derivative Standard Deviation	0.092006
Derivative 75th Percentile	0.078336
Derivative Minimum	0.065031
Derivative Range	0.062746
Derivative 25th Percentile	0.060899
Derivative Maximum	0.059233
Range	0.053030
Signal to Noise Ratio	0.039205
Minimum	0.036135
Standard Deviation	0.035473
Median	0.032935
Maximum	0.032874
50th Percentile	0.031434
25th Percentile	0.030919
Sum of series	0.030824
Mean	0.029730
Geometric Mean	0.029399
Variance	0.029296
75th Percentile	0.029235
Derivative sum of series	0.028591
Derivative Signal to Noise Ratio	0.026762
Derivative Mean	0.026619
Derivative 50th Percentile	0.026240
Derivative Median	0.025393
Derivative Variance	0.007655

Table 4.6: Experiment 3 Extremely Randomized Trees Feature Ranking

It is important to note that the feature ranking is not the same as the random forest, but uses the same method of gini ranking (due to the added randomness introduced in the ERT model).

Iteration 1 involved selecting features which importance is greater than the mean importance of all the features. Using the mean criteria, 7 features were retained: range, derivative minimum, derivative standard deviation, derivative 25th percentile, derivative 75th percentile, derivative maximum, and derivative range. The performance decreased from using all features, so a second iteration was explored.

Iteration 2 selected the features which importance is greater than a chosen threshold of 0.03, retaining 17 features: minimum, standard deviation, mean, 25th percentile, 50th percentile, maximum, range, median, geometric mean, signal to noise ratio, sum of series, derivative minimum, derivative standard deviation, derivative 25th percentile derivative 75th percentile, derivative maximum, and derivative range. The accuracy of the model increased from iteration one, but was still lower than using all features, therefore another iteration was explored.

Iteration 3 selected features with a importance threshold of 0.02, retaining 24 features and rejecting only the derivative variance feature. The accuracy of the model was the best of the three iterations, however was still a lower performance than the ERT model using all 25 features, and was therefore rejected, moving forward with using all available features in the hyperparameter tuning step.

4.8.3 Hyperparameter Tuning

The hyperparameters of the decision tree models that will be focused on in thesis work are number of trees/estimators, maximum depth, minimum samples split, minimum samples leaf, and maximum features. **Number of trees/estimators** refers to the total number of decision trees that make up the entirety of the forest models. **Maximum depth** refers to the maximum

number of levels in each decision tree within the Random Forest. The larger the maximum depth, the deeper the tree, the more splits it has and the greater the ability to capture more information about the data. If set to none, the tree will expand until every leaf is pure; a pure leaf is one in which all data on the leaf comes from the same class. There is no problem with setting the maximum depth of a decision tree model to higher than the number of features. For example, for a model with two features, Age and Sex, one could have a series of splits that first check whether Age > 18, if so check whether Sex = Male, if so check whether Age > 40, etc. Essentially, the same feature can be presented at multiple levels, but with multiple conditions. **Minimum samples split** refers to the number of samples required to split an internal leaf node on a tree. For example, if the minimum samples split is set to two, the decision node must have at least two samples before it can be split to a more specific classification. The **minimum samples leaf** hyperparameter specifies the number of samples required to be at a leaf node. **Maximum features** refers to the number of features to consider when looking for the best split, and cannot be greater than the total number of features within a dataset.

In their paper, Probst et. al., 2017 [77] explores whether or not it is useful to tune the number of trees in a forest classifier in an attempt to improve performance. For random forest and extremely randomized trees models, increasing the number of trees does not result in over-fitting; after a certain threshold of trees, the performance of the classifier does not improve. Over-fitting is not completely avoidable and no model is entirely immune, but the tendency to over-fit decreases as the number of trees increases. As both the RF and ERT classifiers are decision tree-based models, the same hyperparameters were tuned for each, described in Tables 4.7 and 4.8. The tuning ranges were chosen based on [78], with some adjustments made with respect to substituting values appropriate to the thesis work (i.e. number of features equal to the number of features available), and the minimum samples split and leaf criteria ranging from 5% to 100% of the data.

Parameter	Tuning Range
Number of trees (estimators)	1-1000
Maximum Depth	1-31
Minimum Samples Split	9-174
Minimum Samples Leaf	9-174
Maximum Features	1-21

Table 4.7: Hyperaparameter Tuning for RF

Parameter	Tuning Range
Number of trees (estimators)	1-1000
Maximum Depth	1-31
Minimum Samples Split	9-174
Minimum Samples Leaf	9-174
Maximum Features	1-25

Table 4.8: Hyperaparameter Tuning for ERT

Chapter 5

Results

The Results chapter of this thesis includes the performance of the classifiers under the cross-fold evaluation technique, as well as the accuracy performance of the classifiers on the unseen validation dataset.

5.1 Performance of 24-hour time series classification

The classifier architecture used with the 24-hour segmented time series was first Random Forest. Table 5.1 shows the average performance of the classifier using 10-fold cross validation.

Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC
0.5224	0.1755	0.6409	0.5520	0.6071	0.5162

Table 5.1: Average 10-fold cross validation Performance Scores 24-hour period Statistical Features

The accuracy of the classifier using statistical features performed poorly, and on the unseen validation data worse than random guessing with a prediction accuracy of 0.3571. The poor performance led to a re-evaluation of the selected time series window which features were to

be extracted from, and started with manual visualization of the accelerometer data. Chapter 4 includes a detailed explanation of the down-sampling and window selection process used for the second iteration of the time series segmentation for the accelerometer data, and will not be repeated here. Due to the RF model's poor performance, no further classifiers were tested using this data segmentation technique, rather, as mentioned in Section 4, 8-hour time series segmentation was used for the remainder of the thesis work.

5.2 Performance of 8-hour time series classification

5.2.1 Experiment One: Classification using Basic Statistics

The results of experiment one for all four classifiers are shown in Table 5.2.

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.6793	0.1714	0.8303	0.6710	0.6803	0.7887	-0.1900
ERT	0.7007	0.1602	0.8187	0.6844	0.6786	0.7777	-0.1887
LR	0.6204	0.1461	0.7217	0.6342	0.6750	0.7063	-0.2217
SVM	0.6347	0.1710	0.7093	0.6513	0.7143	0.6938	-0.2289

Table 5.2: Average 10-fold cross validation Performance Scores Statistical Features

On the unseen validation set, representing 10% of the total data, the accuracy performance by the RF model was 0.7222, for ERT 0.6666, for LR 0.5000 and for the SVM model 0.6111.

5.2.2 Experiment Two: Classification using Basic Statistics and First Order Derivative

The results of experiment two for all four classifiers are shown in Table 5.3.

On the unseen validation set the RF classifier achieved an accuracy of 0.9444. The ERT accuracy achieved was also 0.9444. Both the LR and SVM models performed much worse, with accuracy performances for LR and SVM 0.5000 and 0.6111, respectively. From examination of experiment 1 and 2's results, as the ERT and RF architectures provided the best performance, these models were chosen to iterate further by tuning model hyperparameters to improve classifier performance in experiment 3.

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.8073	0.0606	0.8806	0.8063	0.8303	0.8749	-0.1397
ERT	0.8207	0.0876	0.8898	0.8261	0.8571	0.8886	-0.1289
LR	0.6204	0.1461	0.7196	0.6342	0.6750	0.7047	-0.2215
SVM	0.6348	0.1710	0.7087	0.6513	0.7143	0.6956	-0.2291

Table 5.3: Average 10-fold cross validation Performance Scores Statistical Features and First-Order Derivative

5.2.3 Experiment 3: Tuning the Hyperparameters of the Decision Tree-based Models

Appendix B contains the 10-fold cross validation curves created for the accuracy, brier score loss, and ROC-AUC of the Random Forest and Extremely Randomized Trees classifiers. These figures represent the changes in the performance of the two models when adjusting the values of the number of trees, maximum depth, minimum samples split, and minimum samples per leaf of the two architectures.

The final parameter settings for the RF and ERT models which yield the highest performance are shown in Tables 5.4 and 5.5. The performance of the models using 10-fold cross validation is shown in Table 5.6 and the confusion matrices for the RF and ERT models on the unseen validation data are found in 5.1 and 5.2. The RF classifier achieved a 0.9444 accuracy on

the 10% validation data, and the ERT classifier also achieved a 0.9444 accuracy on the 10% validation data.

Parameter	Setting
Number of trees (estimators)	300
Split criterion	gini
Max Depth	10
Minimum Samples Split	Default = 2
Minimum Samples Leaf	Default = 1
Maximum Features	4

Table 5.4: Post-Hyperparameter Tuning Model Values for Random Forest

Parameter	Setting
Number of trees (estimators)	50
Split criterion	gini
Max Depth	15
Minimum Samples Split	Default = 2
Minimum Samples Leaf	Default = 1
Maximum Features	16

Table 5.5: Post-Hyperparameter Tuning Model Values for Extremely Randomized Trees

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.8136	0.0643	0.8911	0.8138	0.8428	0.8829	-0.1369
ERT	0.8269	0.0842	0.8903	0.8252	0.8446	0.8888	-0.1298

Table 5.6: Average 10-fold cross validation Performance Scores Experiment Three

Confusion Matrix	Predicted	
	Negative	Positive
Negative	TN = 8	FP = 0
Positive	FN = 1	TP = 9

Figure 5.1: Confusion Matrix for Final RF Model on Set Aside Validation Data

Confusion Matrix	Predicted	
	Negative	Positive
Negative	TN = 8	FP = 0
Positive	FN = 1	TP = 9

Figure 5.2: Confusion Matrix for Final ERT Model on Set Aside Validation Data

Chapter 6

Discussion

ICMEs are complex events that require many pieces of information to correctly identify and characterise. There is no one property of an ICME that is consistent with every event and can be used to definitively describe the phenomena as a whole. Richardson and Cane [24] note that in their ICME database, they "may not have been able to delineate every individual ICME that is present" from 1996-2009, and list two main factors that contribute to this:

- During periods of high solar activity, multiple ICMEs and shocks pass by the Earth which increase the complexity of the interplanetary observations, therefore some ICMEs listed may actually encompass multiple events, and;
- When an extended ICME region has a complicated structure (i.e. complex ejecta), it is difficult to discern how many ICMEs contribute to each storm.

Upon comparison of their catalog with others, the authors noted that the complexity of ICMEs suggest that it is improbable a single parameter can indicate the "true" ICME interval period, and they summarized that ICMEs may exhibit a number of signatures, and the ICME start and end boundaries inferred from these signatures frequently differ. Burlaga, Plunkett, and St. Cyr [79] also note that although multiple CMEs at the Sun may contribute to the production of extended ICMEs, it may be difficult to identify features in the CME that correspond to

individual component CMEs.

The complicated relationship between ICME signatures and identification of storms served as motivation for the completed thesis work; investigating if accelerometer data can be used as a potential new, characteristic signature for ICME identification, space weather monitoring, and solar storm characterisation. From the results presented in Chapter 5, this thesis has been able to demonstrate that satellite accelerometer data utilized by Random Forest and Extremely Randomized Trees binary classifiers can produce accurate results up to 82%. Features can be extracted from accelerometer data to successfully train a classifier to identify ICME events, which supports the claim that accelerometer data can be used as a supplementary characteristic signature to identify ICMEs, without further complex post-processing needed on the data. To be used in atmospheric modelling, the GRACE-A accelerometer data requires that the "measurements cannot be used directly and have to calibrated, as they are affected by an instrument bias and scale" [18]. However, this thesis work has been able to show that, without calibration, a binary classifier can be trained on the accelerometer Level-1B ACC1B data of GRACE-A and perform well in identifying Interplanetary Coronal Mass Ejections from periods of quiet geomagnetic activity.

6.1 Model Performance Evaluation

The decision tree classifiers well outperformed the logistic regression and support vector machine models. Extremely Randomized Trees yielded a slightly higher 10-fold cross validation average accuracy of 0.8269 over the Random Forest accuracy of 0.8136, but had a larger accuracy standard deviation of 0.0842 than the RF's 0.0643. The two models are very comparable with respect to their performance. On the unseen validation data, representing a randomly selected 10% of the total data available (set aside prior to training and testing model with cross validation), both models achieved an accuracy of 0.9444, with 8 true negative predictions, 9

true positive predictions, only one false negative prediction, and zero false positive predictions. Accuracy, accuracy standard deviation, ROC-AUC, and brier loss are the most telling about the classifier performance of the seven performance metrics listed in Table 5.6. Although the ERT model has a higher accuracy (∼1%) than the RF model, the standard deviation of the prediction accuracy is ∼2% larger than the RF model. The average ROC-AUC values are less than a percentage a part, with values of 0.8829 and 0.8888 for the RF and ERT models respectively. The ROC-AUC score gives information as to the classifier's performance in its ability to distinguish between classes, therefore the average ability for the RF and ERT models to distinguish between an 8-hour period containing an ICME disturbed period and an 8-hour geomagnetic quiet period is ∼88%.

Both the Logistic Regression and Support Vector Machine models performed poorly. The Logistic Regression model had an average 10-fold cross validation accuracy of 0.6204, which at first indicates that it could be improved, however the large standard deviation on the accuracy is misleading to the actual performance of the classifier. The unseen validation set classification performance of 0.50 is indicative of the problem with the large standard deviation; 0.50 is within the 0.1461 standard deviation margin of the classifier accuracy, and therefore is not an unexpected result, although on the lower end. It is not entirely unexpected that the LR model would not perform well with the dataset provided, due to the relatively small number of samples available with respect to machine learning datasets, as logistic regression models require fairly large samples for peak performance. To determine the preferred sample size, statistical power, number of parameters to estimate, the percentage of the data containing the "1" class, effect size (strength of the relationship between two variables), and standard error must be considered, which has not been done in this work, due to the lack of flexibility in the available data for the model. Several authors have suggested methods for calculating an appropriate sample size for use in logistic regression modelling, [80] and [81], which could have been used barring data availability. The average accuracy using 10-fold cross validation of the Support Vector Machine classifier was 0.6348, with an accuracy standard deviation of 0.1710, ∼3% higher

than the Logistic Regression model's standard deviation. On the unseen validation data, the SVM model performed at an accuracy of 0.6111. The poor accuracy of the SVM model is somewhat expected, as a linear kernel was chosen for the classifier. A linear kernel assumes the data is linearly separable, this is unlikely in the case of the accelerometer data, however it is the most common kernel to use in a first-iteration, and was therefore pursued. A third-order polynomial kernel was tested with the SVM model, however classification processing times on the order of 10 hours occurred, which is not realistic when considering there is less than 200 samples of 25 features per sample, and the simulations were terminated prior to completion after the 10-hour mark. The Brier Score Loss of SVM and LR were very high, ~22% for both models. The Brier score measures the mean squared difference between the predicted probability assigned to the possible outcomes for a sample, and the actual outcome, and should be minimized. The error for LR and SVM is almost double that of the RF and ERT models, and is another reason why the RF and ERT models were pursued further after experiment three for further iteration to maximize performance.

6.2 False Negatives

An inherent fact of using information derived from one-axis accelerometer data as the only input into a machine learning classifier, along with the relationship between the strength of an interplanetary coronal mass ejection, it's position on the Sun relative to the Earth, and the position of the satellite at the time of the solar storm, is the inevitability of false negatives. A large amount of false negative classifications was found during iteration one (experiment zero) of the down-sampling process discussed in the methodology section. Not all ICMEs are strong enough to cause a significant geomagnetic response which can be captured by the accelerometer, or movement, of the GRACE satellite. Figures 4.5 and 4.6 show the virtually identical behaviour of GRACE-A during a weaker storm and quiet day, prior to setting the criteria that a storm have a Dst index of less than -50 for at least 4 of the 8 total consecutive

hours. The implications with strictly using accelerometer data as a method of space weather identification, is that one could wrongly identify a period in which there is a solar storm caused by an ICME as a period of geomagnetic quiet activity. However, for purposes of satellite orbitography and in-orbit attitude adjustments, a false negative is not necessarily detrimental to the effectiveness of this classifier. From a satellite operation point of view, if the classifier identifies an ICME-related storm, this means that the satellite has moved or been perturbed in some way, which resulted in this event identification. Therefore, the classifier is detecting only those storms strong enough to cause significant physical perturbation to a spacecraft's attitude.

6.3 Classifier Limitations

A distinct limitation of the performance of the RF and ERT classifiers is their inability to classify **all** types of ICME-induced geomagnetic storms. This limitation was shown with the RF classifier's poor performance of identifying 24-hour periods of storm activity prior to enforcing the Dst <-50 threshold in experiment zero. If an ICME does not cause a significant geomagnetic response to the Earth's atmosphere (in this case, defined as a Dst index for at least 4/8 hours <-50), then it is unable to be captured by the GRACE-A Level1B accelerometer data using the current feature set. However, this does not mean that the use of accelerometer data as a characteristic signature for an ICME is not useful. As stated in Richardson & Cane [24], the passage of an ICME may be indicated by various characteristic signatures, as reviewed by Zwickl et al. [27], Gosling [82], Neugebauer and Goldstein [83], and Zurbuchen et al. [4]. Some of these signatures are relatively ubiquitous and observed in many ICMEs, while others are more rare. From this thesis work, by employing the "strong" geomagnetic storm threshold, the accelerometer data was able to yield a good classification accuracy performance. Such results can be useful for identifying ICME storms in particular for satellite operators, whom care more about the storms that will affect their satellites than the storms that won't. From an orbital operation perspective, it does not matter if an ICME occurred if it is not strong enough to

offset the satellite's orbit; an operator cares about the storms that *are strong enough* to cause an orbital perturbation, which is what the tree-based ensemble classifiers described in this thesis do.

Additionally, the data volume available to train the classifier models was a distinct limitation to the performance of the models. A first step to improving a machine learning classifier is increasing the amount of data to train on. The amount of available data from the GRACE-A satellite, following the criteria of ICME events having a Dst value less than -50 for at least 4/8 hours of a 8-hour window, was maximized; i.e. there was no additional data fitting this criteria to use for training. A possible solution to this would be using the GRACE-B data for the same events, thereby essentially doubling the amount of available data for training. However, this strategy was not pursued due to the similarity in behaviour of the GRACE satellites; the two satellites behave almost identically to ICME events and during quiet periods due to their identical geometries and orbit paths, and would therefore not add any "new" information to the classifier.

6.4 Feature Importance

The importance of a feature can be defined as the increase in the classifier's prediction error after the feature's values have been altered. The Random Forest model's top four most important features, as ranked by the gini method in order of descending importance were: Derivative Standard Deviation, Derivative Range, Derivative 75th Percentile, and Derivative Minimum. The Extremely Randomized Trees model had the same top four features although in slightly different order, as ranked in order of descending importance as: Derivative Standard Deviation, Derivative 75th Percentile, Derivative Minimum, Derivative Range. The top four features for both models were extracted from the first-order derivative time series, which verifies the relevancy of utilizing this time series. The standard deviation is a measure of how spread out the

data is from the average. To understand why the standard deviation is ranked highly, consider Figures 6.1 and 6.2, which show the first-order derivative of GRACE-A's accelerometer data during a period of quiet geomagnetic activity and during an ICME period.

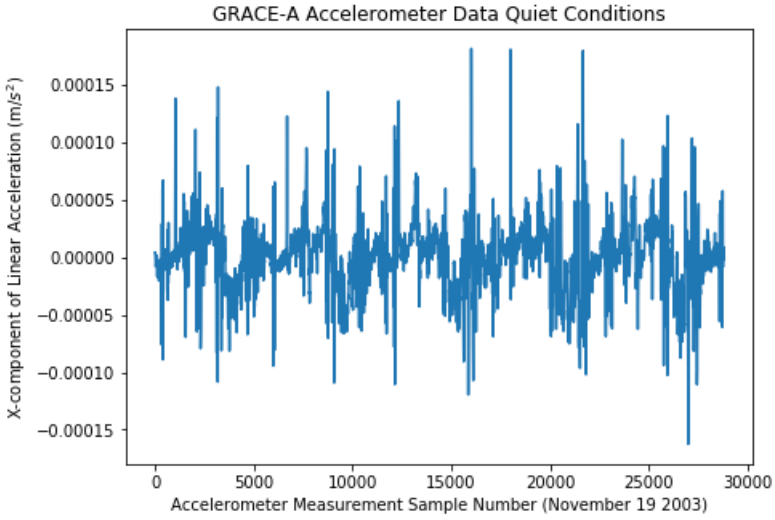

Figure 6.1: First-Order Derivative of GRACE-A Accelerometer data during Quiet Period

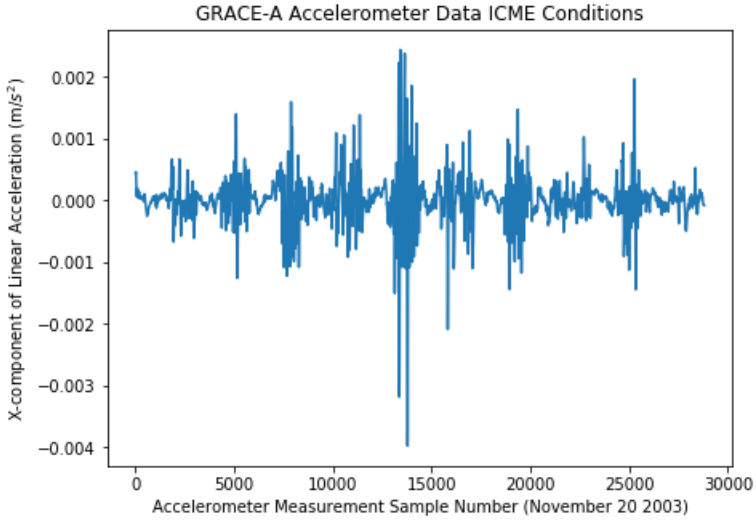

Figure 6.2: First-Order Derivative of GRACE-A Accelerometer data during ICME

Figures 6.1 and 6.2 are examples of data interpreted by the RF and ERT models for the "0" and "1" class, respectively. Visually, the two figures are very different, and the large spread of the time series data around the x-axis at ~15000 of Figure 6.2 is an example of why the standard deviation between the quiet period and ICME period of the time series is an important

feature; the spread of the data from the mean at this point is very obviously different between the two. In fact, the remaining features of the top four; derivative minimum, range, and 75th percentile, all make sense when visually inspecting the derivative time series during a storm and quiet period. The minimum of the ICME period would have a larger magnitude as the difference between a peak in the accelerometer data caused by an ICME to the trough due to the oscillation of the satellite is much greater than the unperturbed behaviour of GRACE-A during a quiet period. The large deviation at ~15000 on the x-axis shown in Figure 6.2 would also correspond to the 75th percentile and range features ranking as higher importance.

Chapter 7

Conclusion

This thesis work presents a new way to utilize machine learning and satellite accelerometer data to identify geomagnetic storms caused by Interplanetary Coronal Mass Ejection phenomena. Previously, space weather modelling has been done using the accelerometer-derived density datasets of the GRACE-A satellite to monitor the changes to the thermosphere due to ICME and flare events. The motivation of this work was to use pre-processed accelerometer data for ICME storm detection, and to investigate the utilization of satellite accelerometer data as a characteristic signature for ICMEs. Utilizing Random Forest, Extremely Randomized Trees, Support Vector Machine, and Logistic Regression models, a maximum binary classification accuracy of 82.69% identification between the time series of geomagnetic storms and geomagnetic quiet periods was achieved using the Extremely Randomized Trees architecture.

7.1 Future Work

An immediate continuation of this research can be done utilizing the CHAMP satellite accelerometer data, whose STAR accelerometer is also used to derive atmospheric density-levels, similar to the GRACE spacecraft. Utilizing CHAMP data would give insight into the robust-

ness of using this structure of classification architecture for identification of ICME-related storms, at a differing satellite altitude (CHAMP's orbit was ~450km while GRACE was launched into ~500km orbit). In addition to utilizing CHAMP, the GRACE-B satellite data could also be used in the exact same way as the GRACE-A data in this thesis, to act as a validation dataset for the proposed methodology.

Furthermore, a future direction to take this research is the implementation of this algorithm on-board a satellite computer, for real-time space weather identification. In addition, the use of this technique with lower-resolution instrumentation could open-up space weather research and monitoring to a wider audience. The GRACE SuperSTAR accelerometer is an expensive, high-resolution, well-designed piece of equipment, specifically designed to accurately measure the Earth's gravitational field. It would be interesting to see the capability of a lower-resolution CubeSat accelerometer for space weather monitoring, to create a cheaper, more accessible way to study ICMEs. Utilizing this technique with other satellites at varying altitudes will increase the overall knowledge of a satellite's response to ICME-related geomagnetic storms, valuable information for satellite operators launching their spacecraft during solar maximum when there is an increase in solar activity.

7.1.1 Machine Learning Techniques

This thesis work provided a proof-of-concept demonstration that satellite accelerometer data can be used to identify space weather events. Future work from a machine learning perspective can include combining the features from the accelerometer with other, well-known ICME signatures (i.e. solar wind, magnetic field measurements, and low proton temperatures). Assessment of the performance of a classifier with these new features will give insight to the relative importance of using accelerometer-derived features to currently existing space weather monitoring techniques.

As was explored throughout this research work, the interval over which the accelerometer data was segmented greatly affected classifier performance (i.e. 24-hour periods that include an ICME storm versus 8-hour periods that include the strongest portion of a storm). Future work related to time series classification with the GRACE dataset could be explored to increase classifier performance, such as employing a time series forest technique as demonstrated by Deng et. al. [14]. Their algorithm uses randomly selected interval periods of an equal length, equally-spaced time series and simple statistical features to achieve a computationally efficient classifier that outperforms one-nearest-neighbour classifiers with dynamic time warping (a commonly used architecture for time series classification). Appendix D includes an excerpt of the results from the author's work, showing it's capabilities for binary classification. It is possible that such a technique could be used with the GRACE dataset, and could be a future direction to take classification of satellite accelerometer data towards.

Bibliography

[1] E. K. Sutton, J. M. Forbes, R. S. Nerem, and T. N. Woods, "Neutral density response to the solar flares of October and November, 2003," *Geophysical Research Letters*, vol. 33, no. 22, pp. 1–5, 2006.

[2] K. Case, G. Kruizinga, and S.-C. Wu, "GRACE Level 1B Data Product User Handbook," tech. rep., 2010.

[3] B. F. V. J. O. P. TOUBOUL, E. WILLEMENOT, "Accelerometers for CHAMP, GRACE and GOCE space missions: synergy and evolution," *BOLLETTINO DIGEOFISICATE-ORICA EDA PPLICATA*, vol. 40, pp. 321–327, 1999.

[4] T. H. Zurbuchen and I. G. Richardson, "In-situ solar wind and magnetic field signatures of interplanetary coronal mass ejections," *Space Science Reviews*, vol. 123, pp. 31–43, 3 2006.

[5] S. Krauss, M. Temmer, and S. Vennerstrom, "Multiple Satellite Analysis of the Earth's Thermosphere and Interplanetary Magnetic Field Variations Due to ICME/CIR Events During 20032015," *Journal of Geophysical Research: Space Physics*, vol. 123, pp. 8884–8894, 10 2018.

[6] "How Decision Tree Algorithm works." `https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/`. Accessed: 2019-07-17.

[7] "Feature Importance Measures for Tree Models Part 1." https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3. Accessed: 2018-06-03.

[8] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, 4 2006.

[9] L. L. D. Jedamski, "Applied Machine Learning: Algorithms."

[10] "Understanding Data Science Classification Metrics in Scikit-Learn in Python." https://towardsdatascience.com/understanding-data-science-classification-metrics-in-scikit-learn-in-python-3b
Accessed: 2019-07-15.

[11] "Understanding AUC - ROC Curve." https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. Accessed: 2019-06-10.

[12] E. Zdravevski, B. R. Stojkoska, M. Standl, and H. Schulz, "Automatic machine-learning based identification of jogging periods from accelerometer measurements of adolescents under field conditions," *PLoS ONE*, vol. 12, 9 2017.

[13] "Train/Test Split and Cross Validation in Python." https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6. Accessed: 2018-08-05.

[14] H. Deng, G. Runger, E. Tuv, and M. Vladimir, "A time series forest for classification and feature extraction," *Information Sciences*, vol. 239, pp. 142–153, 8 2013.

[15] "GRACE Mission Overview," 2012.

[16] B. D. Tapley, S. Bettadpur, M. Watkins, and C. Reigber, "The gravity recovery and climate experiment: Mission overview and early results," *Geophysical Research Letters*, vol. 31,

5 2004.

[17] C. Reiberg, "CHAMP a challenging micro-satellite payload for geophysical research and application," *GFZ Final Report*, 1995.

[18] B. Klinger and T. Mayer-Gürr, "The role of accelerometer data calibration within GRACE gravity field recovery: Results from ITSG-Grace2016," *Advances in Space Research*, vol. 58, pp. 1597–1609, 11 2016.

[19] "Solar Storm and Space Weather - NASA." https://www.nasa.gov/mission_pages/sunearth/spaceweather/index.html. Accessed: 2019-07-25.

[20] "Geomagnetic Storms." https://www.swpc.noaa.gov/phenomena/geomagnetic-storms. Accessed: 2018-10-01.

[21] "Disturbance Storm-Time (Dst) Indices - NCEI." https://www.ngdc.noaa.gov/stp/geomag/dst.html. Accessed: 2019-07-14.

[22] "Dst Index." http://pluto.space.swri.edu/image/glossary/dst.html. Accessed: 2018-05-19.

[23] " Planetary K-index - NOAA / NWS Space Weather Prediction Center." https://www.swpc.noaa.gov/products/planetary-k-index. Accessed: 2019-01-02.

[24] I. G. Richardson and H. V. Cane, "Near-earth interplanetary coronal mass ejections during solar cycle 23 (1996 - 2009): Catalog and summary of properties," *Solar Physics*, vol. 264, no. 1, pp. 189–237, 2010.

[25] R. Schwenn, H. Rosenbauer, and K. . Mühlhäuser, "Singlyionized helium in the driver gas of an interplanetary shock wave," *Geophysical Research Letters*, vol. 7, no. 3, pp. 201–204, 1980.

[26] J. Gosling, J. Asbridge, S. Bame, W. Feldman, and R. Zwickl, " Observations of large fluxes of He + in the solar wind following an interplanetary shock ," *Journal of Geophys-*

ical Research, vol. 85, no. A7, p. 3431, 1980.

[27] R. D. Zwickl, J. R. Asbridge, S. J. Bame, W. C. Feldman, and J. T. Gosling, " He + and other unusual ions in the solar wind: A systematic search covering 19721980 ," *Journal of Geophysical Research*, vol. 87, no. A9, p. 7379, 1982.

[28] Z. G. e. a. Yermolaev Yu.I., Zhuravlev V.I., "Observation of singly-ionized helium in the solar wind," *Journal of Geophysical Research*, vol. 27, no. 5, pp. 717–725, 1989.

[29] L. Burlaga, R. Fitzenreiter, R. Lepping, K. Ogilvie, A. Szabo, A. Lazarus, J. Steinberg, G. Gloeckler, R. Howard, D. Michels, C. Farrugia, R. P. Lin, and D. E. Larson, "A magnetic cloud containing prominence material: January 1997," *Journal of Geophysical Research: Space Physics*, vol. 103, pp. 277–285, 1 1998.

[30] W. C. F. J. T. G. D. J. M. J. T. S. R. L. T. P. R. L. F. B. N. F. N. R. M. Skoug, S. J. Bame and C. W. Smith, "A Prolonged He+ Enhancement Within a Coronal Mass Ejection in the Solar Wind," *Geophysical Research Letters*, vol. 26, pp. 161–164, 1999.

[31] T. F. Lechtenberg, "DERIVATION AND OBSERVABILITY OF UPPER ATMOSPHERIC DENSITY VARIATIONS UTILIZING PRECISION ORBIT EPHEMERIDES," tech. rep.

[32] S. Bruinsma, J. M. Forbes, R. S. Nerem, and X. Zhang, "Thermosphere density response to the 20-21 November 2003 solar and geomagnetic storm from CHAMP and GRACE accelerometer data," *Journal of Geophysical Research: Space Physics*, vol. 111, 6 2006.

[33] Y. Chi, C. Shen, Y. Wang, M. Xu, P. Ye, and S. Wang, "Statistical Study of the Interplanetary Coronal Mass Ejections from 1995 to 2015," *Solar Physics*, vol. 291, pp. 2419–2439, 10 2016.

[34] T. Nieves-Chinchilla, A. Vourlidas, J. C. Raymond, M. G. Linton, N. Al-haddad, N. P. Savani, A. Szabo, and M. A. Hidalgo, "Understanding the Internal Magnetic Field Con-

figurations of ICMEs Using More than 20 Years of Wind Observations," *Solar Physics*, vol. 293, 2 2018.

[35] E. Mitsakou and X. Moussas, "Statistical Study of ICMEs and Their Sheaths During Solar Cycle 23 (1996 - 2008)," *Solar Physics*, vol. 289, no. 8, pp. 3137–3157, 2014.

[36] E. Kilpua, H. E. J. Koskinen, and T. I. Pulkkinen, "Coronal mass ejections and their sheath regions in interplanetary space," *Living Reviews in Solar Physics*, vol. 14, 12 2017.

[37] R. P. Lepping, C. C. Wu, and D. B. Berdichevsky, "Automatic identification of magnetic clouds and cloud-like regions at 1 AU: Occurrence rate and other properties," *Annales Geophysicae*, vol. 23, no. 7, pp. 2687–2704, 2005.

[38] A. Ojeda-Gonzalez, O. Mendes, A. Calzadilla, M. O. Domingues, A. Prestes, and V. Klausner, "An Alternative Method for Identifying Interplanetary Magnetic Cloud Regions," *The Astrophysical Journal*, vol. 837, p. 156, 3 2017.

[39] T. Colak and R. Qahwaji, "Automated solar activity prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares," *Space Weather*, vol. 7, no. 6, 2009.

[40] H. Karimabadi, T. B. Sipes, Y. Wang, B. Lavraud, and A. Roberts, "A new multivariate time series data analysis technique: automated detection of flux transfer events using cluster data," *Journal of Geophysical Research: Space Physics*, vol. 114, 6 2009.

[41] G. W. Prölss, "Magnetic Storm Associated Perturbations of the Upper Atmosphere: Recent Results Obtained by Satellite-Borne Gas Analyzers," tech. rep., 1980.

[42] P. M. Mehta, A. C. Walker, E. K. Sutton, and H. C. Godinez, "New density estimates derived using accelerometers on board the CHAMP and GRACE satellites," *Space Weather*, vol. 15, pp. 558–576, 4 2017.

[43] S. Krauss, M. Temmer, A. Veronig, O. Baur, and H. Lammer, "Thermospheric and geomagnetic responses to interplanetary coronal mass ejections observed by ACE and GRACE: Statistical results," *Journal of Geophysical Research A: Space Physics*, vol. 120, pp. 8848–8860, 10 2015.

[44] C. Xiong, H. Lühr, M. Schmidt, M. Bloßfeld, and S. Rudenko, "An empirical model of the thermospheric mass density derived from CHAMP satellite," *Annales Geophysicae*, vol. 36, pp. 1141–1152, 8 2018.

[45] X. Liu, "CU Scholar The Effects of Composition on Thermosphere Mass Density Response to Geomagnetic Activity," tech. rep., 2013.

[46] "About SWICS 1.1 Level 2 Data." `Srl.caltech.edu`. Accessed: 2019-08-01.

[47] A. B. R. J. A. M. J. W. Y. Boudon, F. Barlier, "Synthesis of flight results of the Cactus accelerometer for accelerations below 109g," *Acta Astronautica*, vol. 6, no. 11, pp. 1387–1398, 1979.

[48] F. A. Marcos and J. M. Forbes, "Thermospheric winds from the satellite electrostatic triaxial accelerometer system," *Journal of Geophysical Research*, vol. 90, no. A7, p. 6543, 1985.

[49] R. W. Z. R. H. T. G. J. C. S. N. N. J. S. P. T. J. S. R. W. S. B. L. W. J. R. M. J. L. H. R. M. H. M. J. J. C. P. B. J. C. M. D. S. G. M. Keating, S. W. Bougher, R. G. W. D. F. R. T. Z. M. D. T. L. P. B. E. M. D. J. C. W. W. C. G. J. R. T. Clancy, R. C. Blanchard, and J. M. Babicke, "The Structure of the Upper Atmosphere of Mars: In Situ Accelerometer Measurements from Mars Global Surveyor," *Science*, vol. 279, no. 5357, pp. 1672–1676, 1998.

[50] S. Bougher', G. Keating2, R. Zurek3, J. Murphy4, R. Haberle5, J. Bo˜ngsworth5, and R. T. Clancy, "MARS GLOBAL SURVEYOR AEROBRAKING : ATMOSPHERIC TRENDS AND MODEL INTERPRETATION," tech. rep., 1999.

[51] C. Berger and F. Barlier, "Response of the equatorial thermosphere to magnetic activity analysed with accelerometer total density data. Asymmetrical structure," *Journal of Atmospheric and Terrestrial Physics*, vol. 43, no. 2, pp. 121–133, 1981.

[52] C. Berger and F. Barlier, "ASYMMETRICAL STRUCTURE IN THE THERMOSPHERE DURING MAGNETIC STORMS AS DEDUCED FROM THE CACTUS ACCELEROMETER DATA," tech. rep.

[53] J. M. Forbes, R. G. Roble, and F. A. Marcos, "Thermospheric dynamics during the March 22, 1979, magnetic storm: 2. Comparisons of model predictions with observations," *Journal of Geophysical Research*, vol. 92, no. A6, p. 6069, 1987.

[54] J. M. Forbes, R. Gonzalez, F. A. Marcos, D. Revelle, and H. Parish, "Magnetic storm response of lower thermosphere density," *Journal of Geophysical Research: Space Physics*, vol. 101, pp. 2313–2319, 2 1996.

[55] J. M. Forbes, G. Lu, S. Bruinsma, S. Nerem, and X. Zhang, "Thermosphere density variations due to the 1524 April 2002 solar events from CHAMP/STAR accelerometer measurements," *Journal of Geophysical Research*, vol. 110, no. A12, 2005.

[56] H. Liu and H. Lühr, "Strong disturbance of the upper thermospheric density due to magnetic storms: CHAMP observations," *Journal of Geophysical Research: Space Physics*, vol. 110, no. A9, 2005.

[57] S. P. Reigber C., Lhr H., *First CHAMP Mission Results for Gravity, Magnetic and Atmospheric Studies*, ch. Bruinsma S., Biancale R., Total density retrieval with STAR. Springer, 2003.

[58] "Introducing SWARM." `https://www.esa.int/Our_Activities/Observing_the_Earth/Swarm/Introducing_Swarm`. Accessed: 2019-06-03.

[59] T. Kodikara, B. Carter, and K. Zhang, "The First Comparison Between Swarm-C Accelerometer-Derived Thermospheric Densities and Physical and Empirical Model Es-

timates," *Journal of Geophysical Research: Space Physics*, vol. 123, pp. 5068–5086, 6 2018.

[60] M. G. Bobra and S. Ilonidis, "PREDICTING CORONAL MASS EJECTIONS USING MACHINE LEARNING METHODS," *The Astrophysical Journal*, vol. 821, p. 127, 4 2016.

[61] E. Camporeale, A. Carè, and J. E. Borovsky, "Classification of Solar Wind With Machine Learning," *Journal of Geophysical Research: Space Physics*, vol. 122, pp. 910–10, 11 2017.

[62] Y. Yang, H. Yang, X. Bai, H. Zhou, S. Feng, and B. Liang, "Automatic detection of sunspots on full-disk solar images using the simulated annealing genetic method," *Publications of the Astronomical Society of the Pacific*, vol. 130, 10 2018.

[63] Y. Wang, D. G. Sibeck, J. Merka, S. A. Boardsen, H. Karimabadi, T. B. Sipes, J. Šafránková, K. Jelínek, and R. Lin, "A new three-dimensional magnetopause model with a support vector regression machine and a large database of multiple spacecraft observations," *Journal of Geophysical Research: Space Physics*, vol. 118, no. 5, pp. 2173–2184, 2013.

[64] X. Miniere, J.-L. Pinçon, and F. Lefeuvre, "A neural network approach to the classification of electron and proton whistlers," *Journal of Atmospheric and Terrestrial Physics*, vol. 58, pp. 911–924, 5 1996.

[65] G. Nguyen, N. Aunai, D. Fontaine, E. L. Pennec, J. V. d. Bossche, A. Jeandet, B. Bakkali, L. Vignoli, and B. R.-S. Blancard, "Automatic Detection of Interplanetary Coronal Mass Ejections from In Situ Data: A Deep Learning Approach," *The Astrophysical Journal*, vol. 874, p. 145, 4 2019.

[66] "Random Forest Simple Explanation." `https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d`. Accessed: 2019-07-18.

[67] "Introduction to Random Forests." https://www.datascience.com/resources/notebooks/random-forest-intro. Accessed: 2019-07-15.

[68] H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *Journal of Korean Academy of Nursing*, vol. 43, no. 2, pp. 154–164, 2013.

[69] "User guide: scikit-learn 0.21.3 documentation." https://scikit-learn.org/stable/user_guide.html#. Accessed: 2019-05-01.

[70] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, Multi-Sensor Based Classification of Daily Life Activities: Comparison with State-of-the-Art Algorithms Using a Benchmark Dataset," *PLoS ONE*, vol. 8, 10 2013.

[71] M. Shoaib, S. Bosch, O. Durmaz Incel, H. Scholten, and P. J. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors (Switzerland)*, vol. 14, pp. 10146–10176, 6 2014.

[72] P. Siirtola and J. Röning, "Recognizing Human Activities User-independently on Smartphones Based on Accelerometer Data," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, p. 38, 10 2012.

[73] S.-C. Wu and G. L. Kruizinga, "Algorithm Theoretical Basis Document for GRACE Level-1B Data Processing," *Document for GRACE Level-1*, 2004.

[74] " WDC for Geomagnetism, Kyoto." http://wdc.kugi.kyoto-u.ac.jp/. Accessed: 2018-08-10.

[75] "Kp-index - Aurora Forecast." http://auroraforecast.is/kp-index/. Accessed: 2019-07-10.

[76] J. Schneider, "Cross Validation." https://www.cs.cmu.edu/~schneide/tut5/node42.html. Accessed: 2019-07-30.

[77] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest?," 5 2017.

[78] M. B. Fraj, "In Depth: Parameter tuning for Random Forest."

[79] L. F. Burlaga, S. P. Plunkett, and O. C. Cyr, "Successive CMEs and complex ejecta," *Journal of Geophysical Research: Space Physics*, vol. 107, no. A10, 2002.

[80] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York, NY: John Wiley & Sons Inc., 2 ed., 1994.

[81] F. Y. Hsieh, D. A. Bloch, and M. D. Larsen, "A SIMPLE METHOD OF SAMPLE SIZE CALCULATION FOR LINEAR AND LOGISTIC REGRESSION," tech. rep., 1998.

[82] J. T. Gosling, "Coronal Mass Ejections and Magnetic Flux Ropes in Interplanetary Space," *Physics of Magnetic Flux Ropes*, vol. 58, 1990.

[83] M. Neugebauer and R. Goldstein, "Particle and Field Signatures of Coronal Mass Ejections in the Solar Wind," *Coronal Mass Ejections*, vol. 1999, 1997.

Appendix A

ICME storms used

Table A.1: ICME 8-hour periods used

Storm Start	Storm End	8hr High Dst Period	Max Dst
2002/08/02 0600	2002/08/04 0200	2002/08/02 0600 - 2002/08/02 1400	-85
2002/08/19 1200	2002/08/21 1400	2002/08/21 0000 - 2002/08/21 0800	-106
2002/09/07 1200	2002/09/08 0400	2002/09/07 2000 - 2002/09/08 0400	-181
2002/09/08 0400	2002/09/08 2000	2002/09/08 0400 - 2002/09/08 1200	-149
2002/09/08 2200	2002/09/10 2100	2002/09/08 2200 - 2002/09/09 0600	-79
2002/09/30 2000	2002/10/01 1500	2002/10/01 0700 - 2002/10/01 1500	-158
2002/10/03 0100	2002/10/04 1800	2002/10/04 0500 - 2002/10/04 1300	-146
2002/11/17 1000	2002/11/19 1200	2002/11/18 1900 - 2002/11/19 0300	-51
2002/12/21 0300	2002/12/22 1900	2002/12/21 0300 - 2002/12/21 1100	-75
2003/02/01 1900	2003/02/03 0700	2003/02/02 1400 - 2003/02/02 2200	-72
2003/03/20 1200	2003/03/20 2200	2003/03/20 1400 - 2003/03/20 2200	-64
2003/05/09 0700	2003/05/11 0000	2003/05/10 0300 - 2003/05/10 1100	-84
2003/05/30 0200	2003/05/30 1600	2003/05/30 0200 - 2003/05/30 1000	-135

Continued on next page

Table A.1 – *Continued from previous page*

Storm Start	Storm End	8hr High Dst Period	Max Dst
2003/05/30 2200	2003/06/01 0100	2003/05/30 2300 - 2003/05/31 0700	-63
2003/06/15 2000	2003/06/16 2100	2003/06/16 1200 - 2003/06/16 2000	-59
2003/06/17 1000	2003/06/18 0800	2003/06/17 1000 - 2003/06/17 1800	-77
2003/08/18 0100	2003/08/19 1500	2003/08/18 1400 - 2003/08/18 2200	-148
2003/10/22 0200	2003/10/24 1500	2003/10/22 0200 - 2003/10/22 1000	-61
2003/10/29 1100	2003/10/30 0300	2003/10/29 1900 - 2003/10/30 0300	-353
2003/10/31 0200	2003/11/02 0000	2003/10/31 0200 - 2003/10/31 1000	-244
2003/11/20 1000	2003/11/21 0800	2003/11/20 1600 - 2003/11/21 0000	-422
2004/01/22 0800	2004/01/23 1700	2004/01/22 1200 - 2004/01/22 2000	-130
2004/04/03 1400	2004/04/05 1800	2004/04/03 1900 - 2004/04/04 0300	-117
2004/07/22 1800	2004/07/24 0800	2004/07/22 2300 - 2004/07/23 0700	-99
2004/07/24 1400	2004/07/25 1500	2004/07/25 0600 - 2004/07/25 1400	-126
2004/07/25 2000	2004/07/26 2200	2004/07/25 2000 - 2004/07/26 0400	-122
2004/07/27 0200	2004/07/27 2200	2004/07/27 1000 - 2004/07/27 1800	-170
2004/08/29 1900	2004/08/30 2200	2004/08/30 1400 - 2004/08/30 2200	-117
2004/11/07 2200	2004/11/09 1000	2004/11/08 0200 - 2004/11/08 1000	-374
2004/11/09 2000	2004/11/11 2300	2004/11/10 0600 - 2004/11/10 1400	-263
2004/11/12 0800	2004/11/13 2300	2004/11/12 0900 - 2004/11/12 1700	-92
2004/12/12 2200	2004/12/13 1900	2004/12/13 0100 - 2004/12/13 0900	-56
2005/01/07 1500	2005/01/08 1200	2005/01/07 2300 - 2005/01/08 0700	-93
2005/01/16 1400	2005/01/17 0700	2005/01/16 2200 - 2005/01/17 0700	-65
2005/01/18 2300	2005/01/20 0300	2005/01/19 1000 - 2005/01/19 1800	-80
2005/01/21 1900	2005/01/22 1700	2005/01/22 0100 - 2005/01/22 0900	-97

Continued on next page

Table A.1 – *Continued from previous page*

Storm Start	Storm End	8hr High Dst Period	Max Dst
2005/05/15 0600	2005/05/19 0000	2005/05/15 0700 - 2005/05/15 1500	-247
2005/05/20 0300	2005/05/22 0200	2005/05/21 0500 - 2005/05/21 1300	-63
2005/05/30 0100	2005/05/30 2300	2005/05/30 1200 - 2005/05/30 2000	-113
2005/05/31 0400	2005/06/01 0300	2005/05/31 0400 - 2005/05/31 1200	-71
2005/06/12 1500	2005/06/13 1300	2005/06/12 2000 - 2005/06/13 0400	-106
2005/07/10 1000	2005/07/12 0400	2005/07/10 1900 - 2005/07/11 0300	-92
2005/07/17 1400	2005/07/18 2300	2005/07/18 0200 - 2005/07/18 1000	-67
2005/08/24 1400	2005/08/24 2300	2005/08/24 1400 - 2005/08/24 2200	-145
2005/09/11 0500	2005/09/12 0700	2005/09/11 0700 - 2005/09/11 1500	-139
2005/09/12 2000	2005/09/13 1300	2005/09/12 2000 - 2005/09/13 0400	-89
2005/09/13 1600	2005/09/14 0800	2005/09/13 1600 - 2005/09/14 0000	-78
2005/09/15 0600	2005/09/16 1800	2005/09/15 1900 - 2005/09/16 0300	-76
2006/04/14 1300	2006/04/14 2100	2006/04/14 1300 - 2006/04/14 2100	-81
2006/11/29 0500	2006/11/30 1000	2006/11/30 0200 - 2006/11/30 1000	-66
2006/12/14 2200	2006/12/15 1300	2006/12/15 0100 - 2006/12/15 0900	-162
2006/12/15 2000	2006/12/16 1900	2006/12/15 2000 - 2006/12/16 0400	-81
2010/04/05 1200	2010/04/06 1400	2010/04/06 0600 - 2010/04/06 1400	-79
2010/05/28 1900	2010/05/29 1700	2010/05/29 0900 - 2010/05/29 1700	-80
2010/08/04 1000	2010/08/05 0000	2010/08/04 1600 - 2010/08/05 0000	-74
2011/05/28 0500	2011/05/28 2100	2011/05/28 1000 - 2011/05/28 1800	-80
2011/08/06 2200	2011/08/07 2200	2011/08/06 2200 - 2011/08/07 0600	-54
2011/09/10 0300	2011/09/10 1500	2011/09/10 0300 - 2011/09/10 1100	-75
2011/09/17 1400	2011/09/18 0600	2011/09/17 1400 - 2011/09/17 2200	-72

Continued on next page

Table A.1 – *Continued from previous page*

Storm Start	Storm End	8hr High Dst Period	Max Dst
2011/09/26 2000	2011/09/28 1500	2011/09/26 2000 - 2011/09/27 0400	-118
2011/10/24 2200	2011/10/25 1600	2011/10/25 0000 - 2011/10/25 0800	-147
2011/11/02 0100	2011/11/03 0400	2011/11/02 0800 - 2011/11/02 1600	-58
2012/01/22 2300	2012/01/23 0700	2012/01/22 2300 - 2012/01/23 0700	-71
2012/02/14 2100	2012/02/16 0600	2012/02/15 1100 - 2012/02/15 1900	-67
2012/03/09 0300	2012/03/11 0700	2012/03/09 0800 - 2012/03/09 1600	-145
2012/03/15 1700	2012/03/16 1000	2012/03/15 1800 - 2012/03/16 0200	-88
2013/01/17 1600	2013/01/18 1200	2013/01/17 1800 - 2013/01/18 0200	-52
2013/06/28 0200	2013/06/29 1200	2013/06/29 0100 - 2013/06/29 0900	-102
2013/07/05 0100	2013/07/07 1600	2013/07/06 1400 - 2013/07/06 2200	-87
2013/07/13 0500	2013/07/15 0000	2013/07/14 1600 - 2013/07/15 0000	-81
2014/02/18 1500	2014/02/19 0700	2014/02/18 2300 - 2014/02/19 0700	-97
2014/02/19 1200	2014/02/20 0300	2014/02/19 1200 - 2014/02/19 2000	-71
2014/02/21 0200	2014/02/22 1200	2014/02/21 2300 - 2014/02/22 0700	-64
2014/04/11 0600	2014/04/12 2000	2014/04/12 0500 - 2014/04/12 1300	-87
2014/04/29 2000	2014/04/30 2100	2014/04/30 0700 - 2014/04/30 1500	-67
2014/09/12 2200	2014/09/14 0200	2014/09/12 2200 - 2014/09/13 0600	-88
2015/03/17 1300	2015/03/18 0500	2015/03/17 1900 - 2015/03/18 0300	-222
2015/04/10 1300	2015/04/11 0900	2015/04/11 0100 - 2015/04/11 0900	-73
2015/06/23 0200	2015/06/24 1400	2015/06/23 0300 - 2015/06/23 1100	-204
2015/06/25 1000	2015/06/26 0600	2015/06/25 1400 - 2015/06/25 2200	-86
2015/06/26 1200	2015/06/27 0200	2015/06/26 1300 - 2015/06/26 2100	-57
2015/07/13 0600	2015/07/14 1500	2015/07/13 1000 - 2015/07/13 1800	-61

Continued on next page

Table A.1 – *Continued from previous page*

Storm Start	**Storm End**	**8hr High Dst Period**	**Max Dst**
2015/08/15 2100	2015/08/16 0800	2015/08/16 0000 - 2015/08/16 0800	-84
2015/08/26 0800	2015/08/28 1000	2015/08/27 1600 - 2015/08/28 0000	-92
2015/09/08 0000	2015/09/09 1500	2015/09/09 0800 - 2015/09/09 1600	-98
2015/12/20 0300	2015/12/21 2000	2015/12/20 2100 - 2015/12/21 0500	-155
2015/12/31 1700	2016/01/02 1100	2015/12/31 2300 - 2016/01/01 0700	-110
2016/01/19 1000	2016/01/21 0000	2016/01/20 1400 - 2016/01/20 2200	-93
2017/05/27 2200	2017/05/29 1400	2017/05/28 0300 - 2017/05/28 1100	-125

Appendix B

Experiment Three Validation Curves

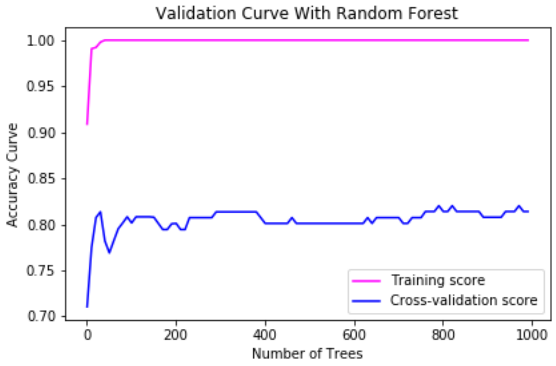

Figure B.1: Random Forest Validation Accuracy Curve Iterating Number of Trees

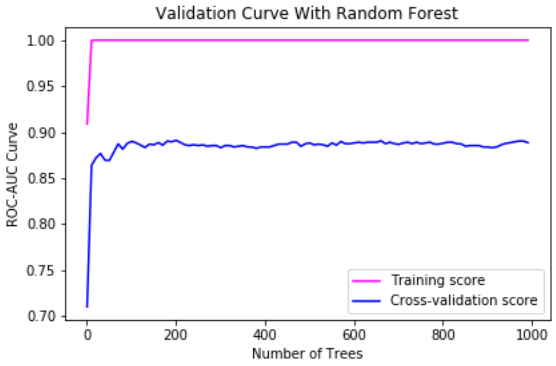

Figure B.2: Random Forest Validation ROC-AUC Curve Iterating Number of Trees

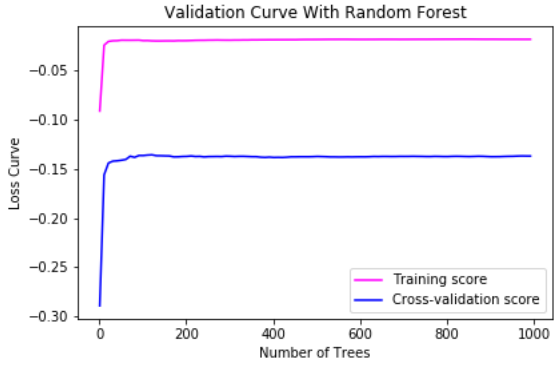

Figure B.3: Random Forest Validation Brier Score Loss Curve Iterating Number of Trees

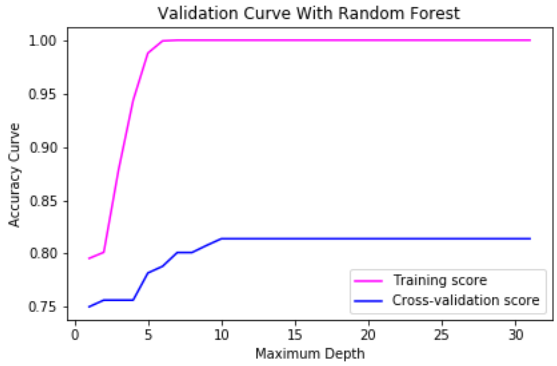

Figure B.4: Random Forest Validation Accuracy Curve Iterating Max Depth

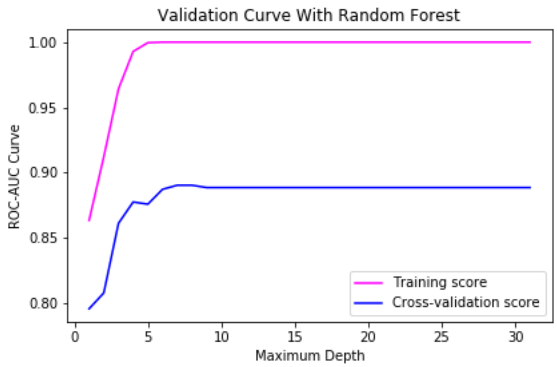

Figure B.5: Random Forest Validation ROC-AUC Curve Iterating Max Depth

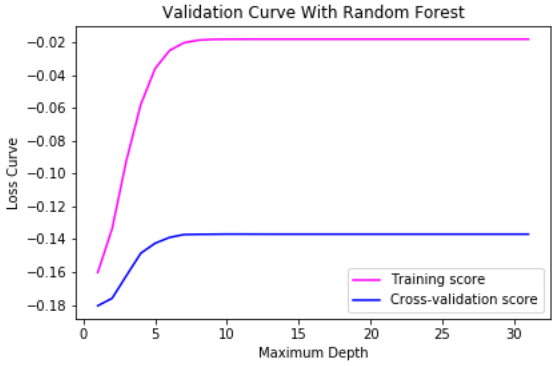

Figure B.6: Random Forest Validation Brier Score Loss Curve Iterating Max Depth

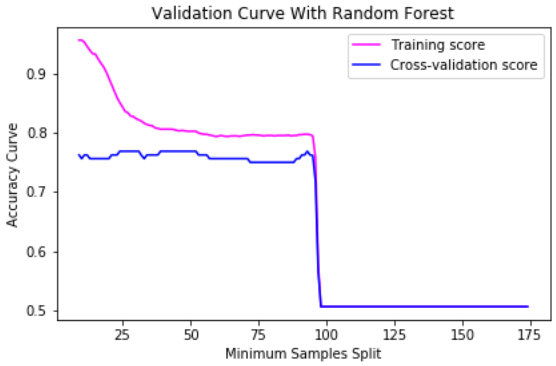

Figure B.7: Random Forest Validation Accuracy Curve Iterating Minimum Samples Split

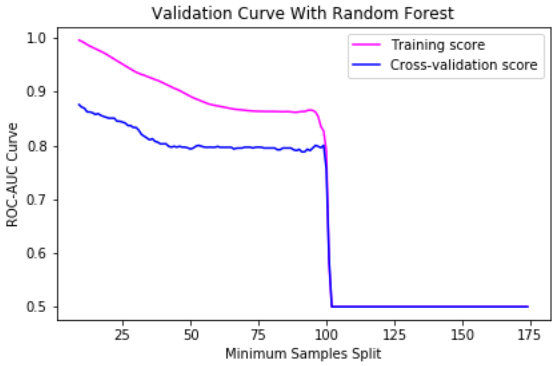

Figure B.8: Random Forest Validation ROC-AUC Curve Iterating Minimum Samples Split

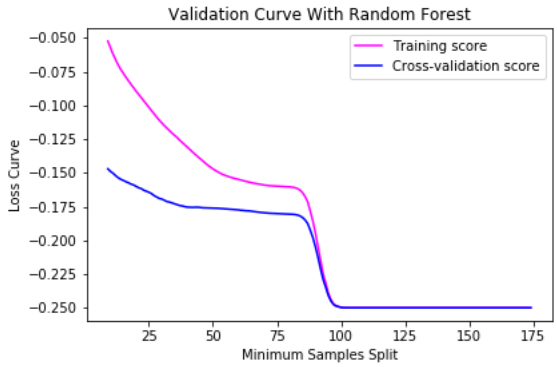

Figure B.9: Random Forest Validation Brier Score Loss Curve Iterating Minimum Samples Split

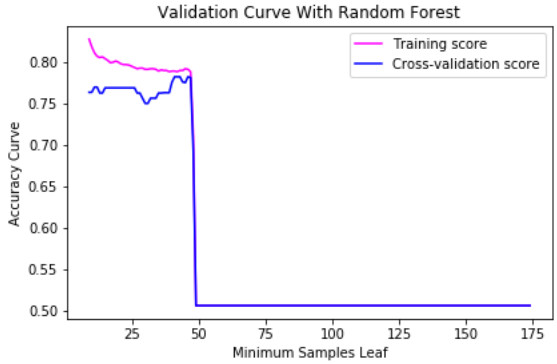

Figure B.10: Random Forest Validation Accuracy Curve Iterating Minimum Samples Leaf

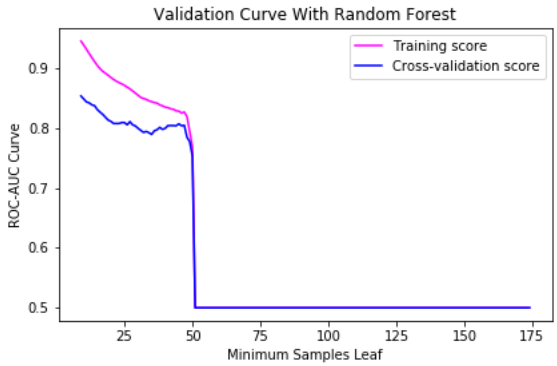

Figure B.11: Random Forest Validation ROC-AUC Curve Iterating Minimum Samples Leaf

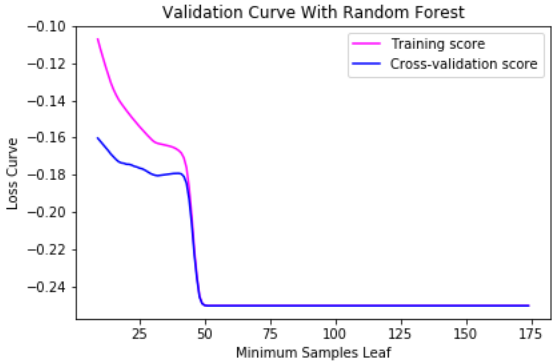

Figure B.12: Random Forest Validation Brier Score Loss Curve Iterating Minimum Samples Leaf

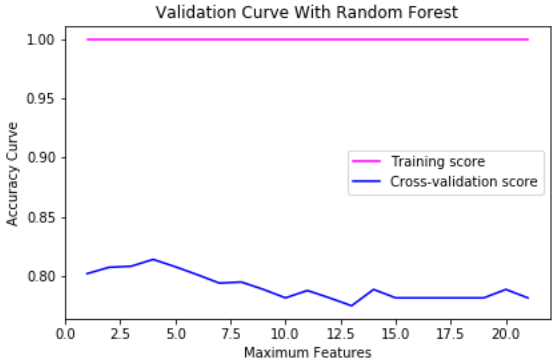

Figure B.13: Random Forest Validation Accuracy Curve Iterating Maximum Features

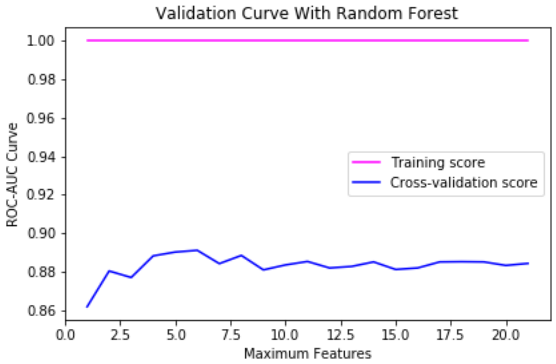

Figure B.14: Random Forest Validation ROC-AUC Curve Iterating Maximum Features

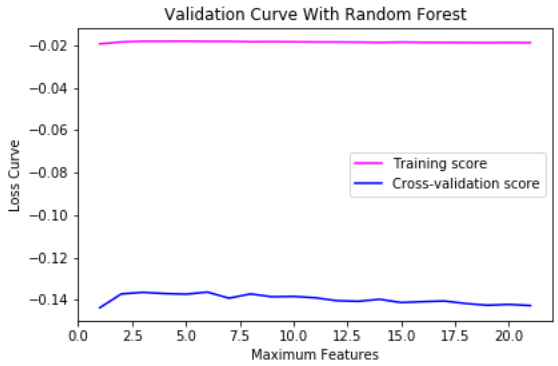

Figure B.15: Random Forest Validation Brier Score Loss Curve Iterating Maximum Features

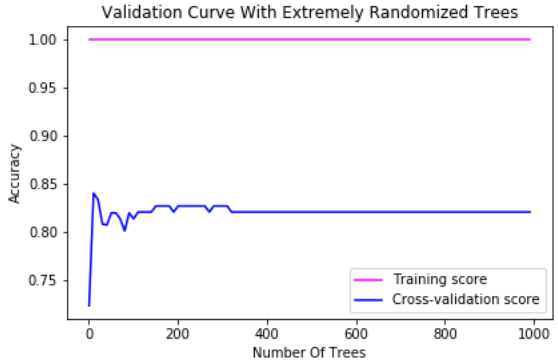

Figure B.16: Extremely Randomized Trees Validation Accuracy Curve Iterating Number of Trees

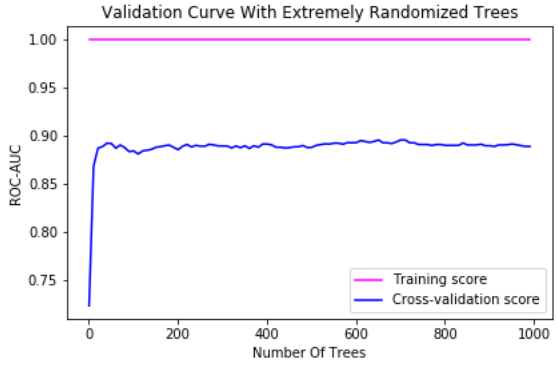

Figure B.17: Extremely Randomized Trees Validation ROC-AUC Curve Iterating Number of Trees

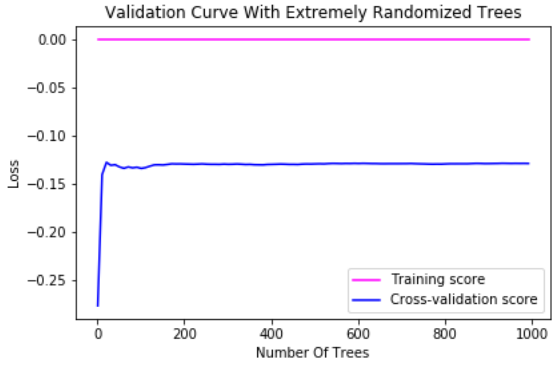

Figure B.18: Extremely Randomized Validation Brier Score Loss Curve Iterating Number of Trees

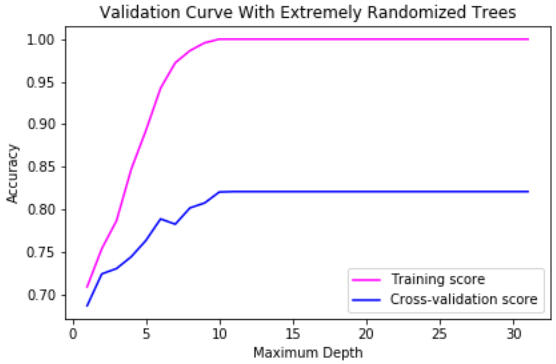

Figure B.19: Extremely Randomized Trees Validation Accuracy Curve Iterating Maximum Depth

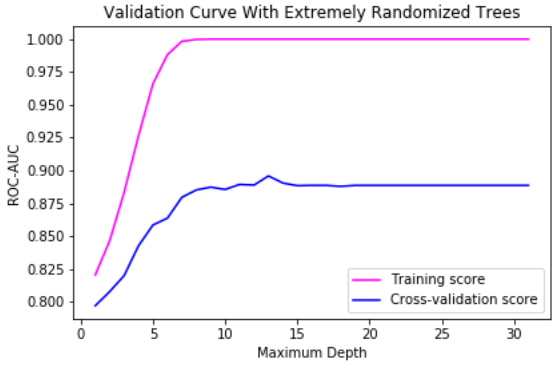

Figure B.20: Extremely Randomized Trees Validation ROC-AUC Curve Iterating Maximum Depth

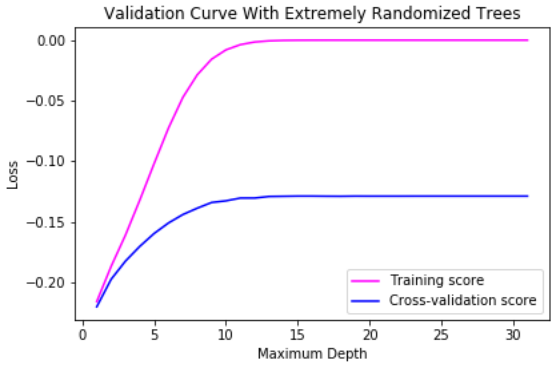

Figure B.21: Extremely Randomized Validation Brier Score Loss Curve Iterating Maximum Depth

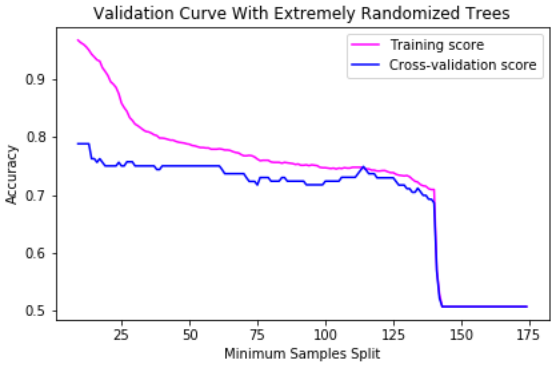

Figure B.22: Extremely Randomized Trees Validation Accuracy Curve Iterating Minimum Samples Split

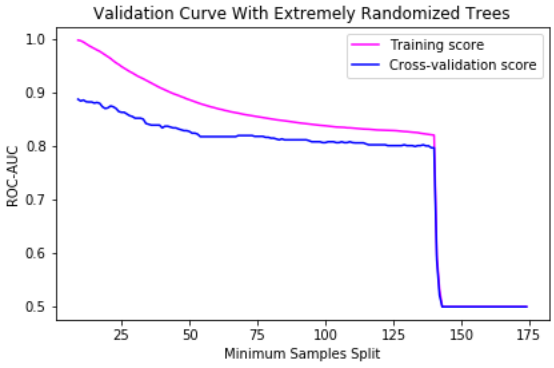

Figure B.23: Extremely Randomized Trees Validation ROC-AUC Curve Iterating Minimum Samples Split

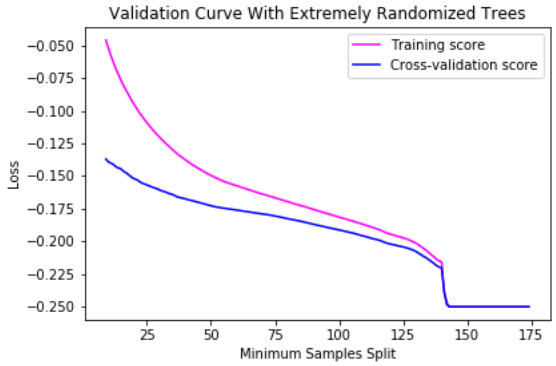

Figure B.24: Extremely Randomized Validation Brier Score Loss Curve Iterating Minimum Samples Split

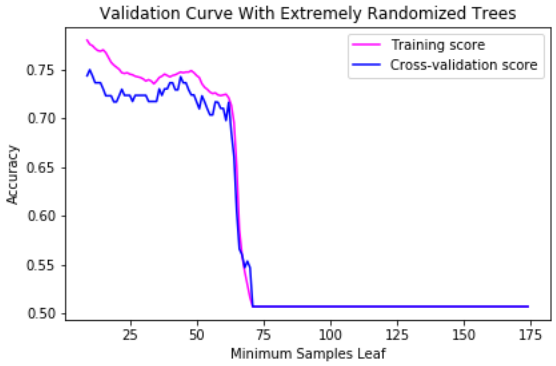

Figure B.25: Extremely Randomized Trees Validation Accuracy Curve Iterating Minimum Samples Leaf

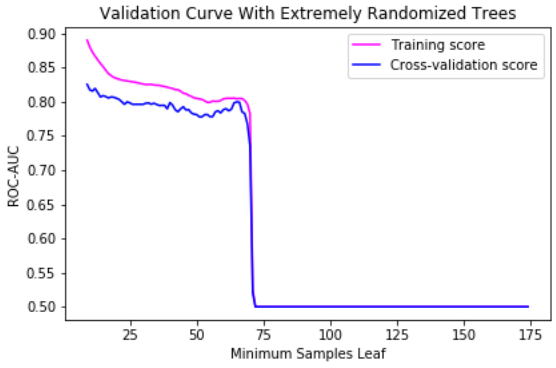

Figure B.26: Extremely Randomized Trees Validation ROC-AUC Curve Iterating Minimum Samples Leaf

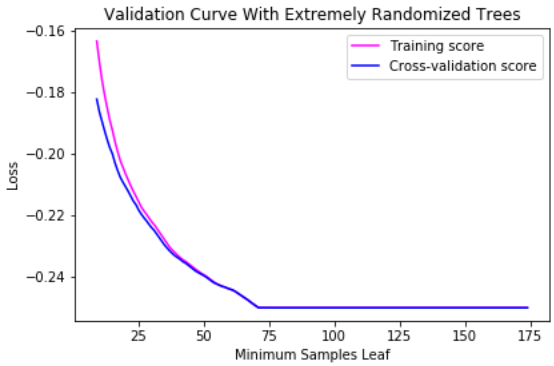

Figure B.27: Extremely Randomized Validation Brier Score Loss Curve Iterating Minimum Samples Leaf

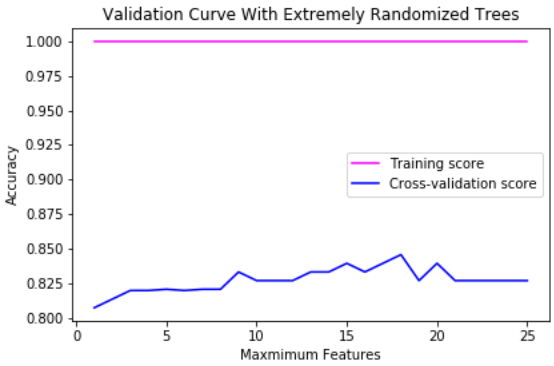

Figure B.28: Extremely Randomized Trees Validation Accuracy Curve Iterating Maximum Features

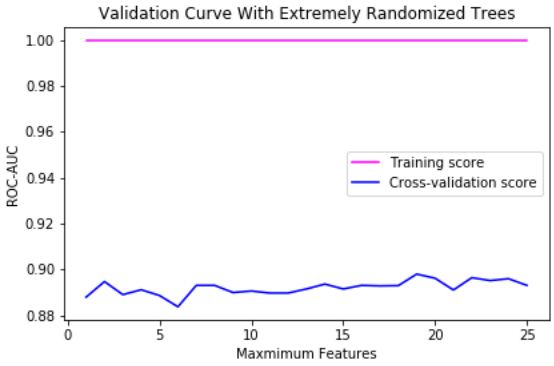

Figure B.29: Extremely Randomized Trees Validation ROC-AUC Curve Iterating Maximum Features

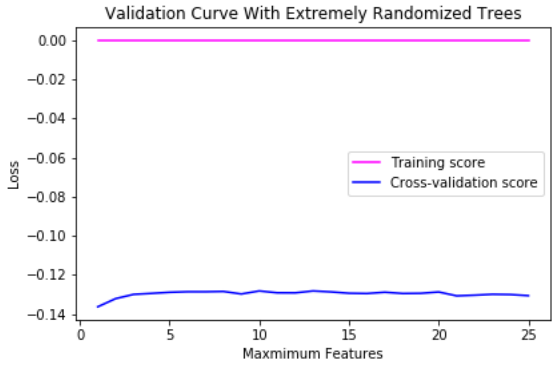

Figure B.30: Extremely Randomized Validation Brier Score Loss Curve Iterating Maximum Features

Appendix C

Experiment Three Iteration 1-3 Results

C.1 Random Forest

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.7502	0.0873	0.8548	0.7425	0.7553	0.8616	-0.1649

Table C.1: Random Forest Experiment 3: Iteration 1 Model Performance

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.7694	0.0797	0.8825	0.7649	0.7928	0.8906	-0.1437

Table C.2: Random Forest Experiment 3: Iteration 2 Model Performance

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.8140	0.0707	0.8965	0.8121	0.8303	0.8884	-0.1369

Table C.3: Random Forest Experiment 3: Iteration 3 Model Performance

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
Random Forest	0.7881	0.0747	0.8938	0.7905	0.8178	0.8974	-0.1378

Table C.4: Random Forest Experiment 3: Iteration 4 Model Performance

C.2 Extremely Randomized Trees

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
ERT	0.7569	0.0898	0.8522	0.7596	0.7946	0.8566	-0.1625

Table C.5: Extremely Randomized Trees Experiment 3: Iteration 1 Model Performance

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
ERT	0.8145	0.0855	0.9054	0.8137	0.8446	0.9006	-0.1267

Table C.6: Extremely Randomized Trees Experiment 3: Iteration 2 Model Performance

Model	Accuracy	Accuracy σ	Precision	F1	Recall	ROC-AUC	Brier Loss
ERT	0.8198	0.0941	0.8904	0.8239	0.8553	0.8895	-0.1288

Table C.7: Extremely Randomized Trees Experiment 3: Iteration 3 Model Performance

Appendix D

Future Work Directions

Table D.1: H.,Deng et al's Classifier Performance [14]

Data Set Used	Classifier Error Rate using Author's Proposed Method
Coffee	0.0357
ECG200	0.0800
ECGFiveDays	0.0800
GunPoint	0.0467
ItalyPowerDemand	0.0301
Lighting2	0.1803
MoteStrain	0.1190
SonyAIBORobotSurface	0.2330
SonyAIBORobotSurfaceII	0.1868
TwoLeadECG	0.1177
Wafer	0.0054
Yoga	0.1513

Curriculum Vitae

Name:	Kelsey Doerksen
Post-Secondary Education and Degrees:	Carleton University Ottawa, ON 2013-2017 B.Eng
	University of Western Ontario London, ON 2018 - 2019 MESc
Honours and Awards:	Outstanding Research Symposium Presentation 2019
	Victor Hangan Global Opportunities Award 2018
Related Work Experience:	JPL Visiting Research Internship NASA Jet Propulsion Laboratory 2019
	Research Internship l'Observatoire de Paris 2018
	Teaching Assistant University of Western Ontario 2018-2019
	Research Assistant University of Western Ontario 2018-2019