National College of Ireland

Data Driven Approach to Predict Solar Flares

MSc Research Project

MSc In Data Analytics

Akshay Ganesh Bhujbal

Student ID: x15041433

School of Computing

National College of Ireland

Supervisor: Dr. Sachin Sharma

Data Driven Approach to Predict Solar Flares

National College of Ireland

Akshay Bhujbal

Student ID x15041433

Abstract

In lay man term solar flare can be defined as an event where there is a sudden flash of amplified brightness observed on the Sun generally close to sunspot group. It is estimated that a flare releases enough energy to power the United States of America for 1 million years. Solar flare has the potential to affect all the layers of atmosphere which include the photosphere, chromosphere and corona. The technologies stationed on the land, space, the astronauts, the aircrew, passengers etc all face the adverse impact of the storm caused by solar radiation. These solar radiations are the by-product of the Solar flares. Hence accurate forecasting of extra-terrestrial weather conditions is essential to mitigate the adversities caused by the unconducive weather conditions. The detection and prediction of spiking solar flares like the X-class and the M-class flare can help save copious amount of resources and capital. Hence with the help of data-driven approaches, concepts of machine learning and our novel approach of amalgamating modified feature extractor with autoregressive model, deep learning sequence neural network and ARIMA model respectively, we can predict precisely the existence of the next Solar Flare by pulling out insights from the extensive time series data with the aim to mitigate the risk induced by ensuring proper contingency plan.

1 Introduction

Solar flares seem fascinating and far away event but have potential to damage the terrestrial satellite and the far away ground based technology. In lay man term solar flare can be defined as an event where there is a sudden flash of amplified brightness observed on the Sun generally close to sunspot group. A solar flare is an explosion on the surface of the Sun which ranges from minutes to hours in length of existence and happen when powerful magnetic field in and around the Sun reconnects which is associated with active region where the magnetic field is strongest. It is calculated that a flare releases enough energy to power the united states of America for 1 million years.

Solar flare has the potential to affect all the layers of atmosphere which include the photosphere, chromosphere and corona. Solar flare emits electromagnetic radiation of all wavelength across the entire electromagnetic spectrum. The excited flux particles have the potential to cause space weather disruption. The technologies stationed on the land, space, the astronauts, the aircrew, passengers etc all face the adverse impact of the storm caused by solar radiation. These solar radiations are the buy product of the Solar flares. Hence a precise and

accurate forecasting of extra-terrestrial weather conditions is essential to mitigate the adversities caused by the unconducive weather conditions. The detection, prediction of spiking solar flares like the X-class and the M-class flare can help save copious amount of resources and capital (Eastwood *et al.*, 2017).

The former dignitaries of United States have filed a plea to the space scientists to dedicate their prime attention to figure out conducive contingency plan to mitigate space weather activities (Eastwood *et al.*, 2017) . The Coronal Mass Ejections and the Solar Flares directly impacts the Earth's magnetic field which has the potential to instigate geomagnetic storms. Historically earth has experienced events like "Carrington Event", blackout in Quebec causing a massive power grid failure in year 1859 and 1989 respectively (Townsend *et al.*, 2006). A solar storm was recorded in the year 2012 causing a hazardous geomagnetic storm underlining the delicacy of the situation and its vulnerability (Townsend *et al.*, 2006).

The main problem of faced by the heliophysics community is that the existing theoretical models which are responsible for the detection and prediction of the Solar flare are not able to explain the alleged relationship of coronal magnetic field and photosphere. Therefore, the entire heliophysics community have pivoted their approach to models with data-driven approaches for the prediction of flair (Hamdi *et al.*, 2018). Generally, a flare occurrence is associated with Active regions of the Sun, therefore a Solar Flare prediction can be modelled in the category of the supervised learning in the field of machine learning. Further this problem can be classed as a classic binary classification where the Active Regions of the Sun are tagged as the positive classes or flaring, and the non-flaring Active regions are strictly classified as non-flaring Active-Regions or the negative class.

The rest of the research proposal will constitute of the Literature review, Research Gap, Research Question, Proposed Methodology, Implementation, Evaluation, Conclusion, Project Plan Gantt chat and References.

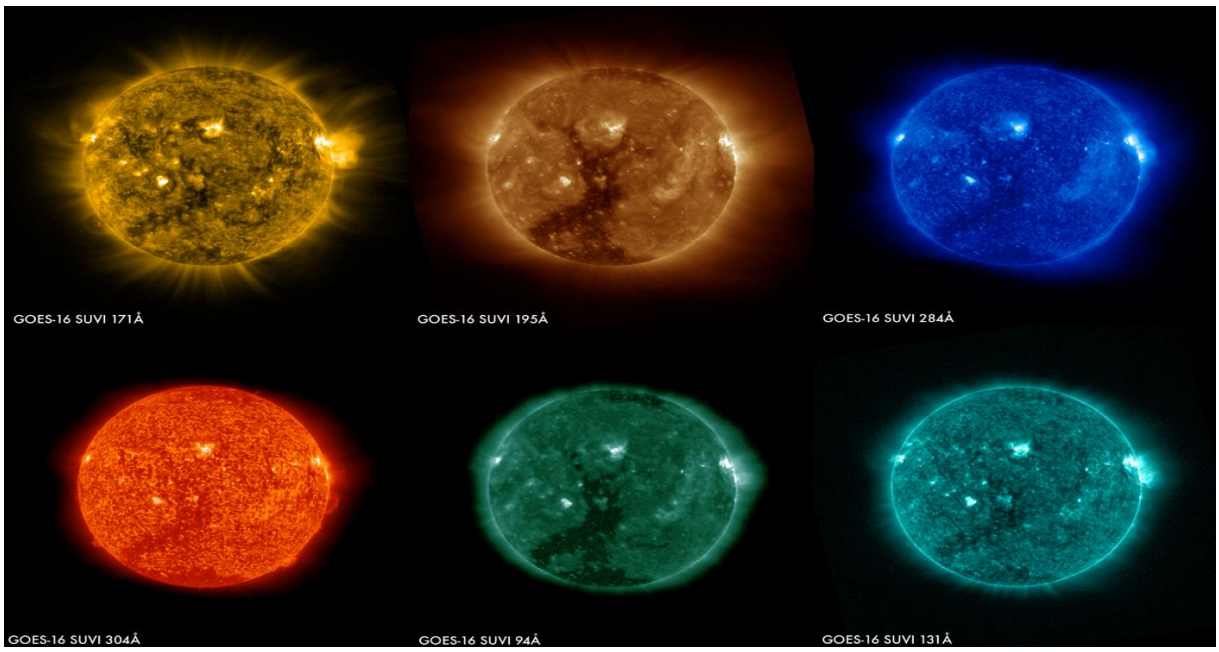


Figure 1: Solar Flares Captured by GOES-17.

2 Related Work

The primary reason to perform feature selection is to pre-optimise the classifier. There is an established trade-off between subset dataset size and the classification performance. Feature selection first decrease the computation cost by decreasing the size of dataset by selecting only relevant features. Second it helps to generate insights by providing the mathematical information and reasons which makes a feature relevant, these insights are descriptive to humans. Feature selection is categorized under problem of grid searching hence recursively search feature in a dataset based on mathematically designed measures and rules (Alto, 1997) . Feature selection methods are subdivided into category of filter, wrapper, or embedded methods depending upon the classification and selection criteria.

Filter methods is based on principle of implementing a selection rule prior to invoke a dedicated classification protocol. The advantage of this rule is ease of applicability and computation efficiency. The independent relation between the filter and classifier leads to average performance of the selection procedure as compared to wrapper and embedded methods.

On the other hand, the wrapper method executes a classification process as primary criteria for feature subset search. The performance of wrapper method is better than the filter method because of the dependent classification method embedded in wrappers. The trade-off of the wrapper method is the computational complexity and the returned feature set is more classifier specific than dataset specific.

The most superior kind of feature selection method is the Embedded method where feature selection executes as process of optimization of classifier. The subset of selected features and the classifier are simultaneously optimised. Some examples of embedded methods are SVM (Guyon *et al.*, 2012) and RVM. The embedded procedures are not implemented in an automated ML toolbox.

1.1 Existing Research in the Solar flare Prediction

As mentioned in the introduction that the approach to predict the existence of solar fare is pivoted to data-driven approaches (Hamdi *et al.*, 2018) the pre-existing flare system to predict flare was THEO (McIntosh, 1990) , a highly efficient system but required human involvement for functionality. The system employed amalgamation of sunspot and properties of the magnetic field for the prediction of the various classes of the Solar flares. Post the purely theoretical modelling the prediction approach was pivoted to data-monitored approaches. The approaches under the umbrella of data-monitored approaches are further classified into two categories linear and non-linear statistics. These established concepts were further classified into a line of sight magnetogram-based models (LSMBM) and vector magnetogram-based models (VMBM).

In 2010 NASA launch an initiative known as Solar Dynamics Observatory (SDO), with an objective to map the vector-magnetic field in every interval of 12 minutes (Mason and Hoeksema, 2010). With advancement in technologies it is established that the data in form of the continuous stream of vector magnetogram work precisely to monitor the Active Regions.

Post 2010 by availability of continuous stream vector magnetogram the researchers abandon the usage of line of sight magnetic data for predicting the flare.

The epicentre of the Linear statistical models in this domain is the identification of the Active Region magnetic properties which are correlated with existence or prediction of the flares. (Cui *et al.*, 2007) and (Jing *et al.*, 2006) employed the use of line of sight magnetogram to characterise the existing of the Active Regions and researched the statistical relationship between the occurrence of the flares and Active Region parameters. (Leka and Barnes, 2003b) were first researchers to exploit the vector magnetogram based approach for Active Region parameters and supplemented the classification process using linear discriminant analysis (LDA). During the summit of Mount Haleakal the data for these researches was archived from Mees Solar Observatory Imaging Vector Magnetograph.

All the Nonlinear statistical models are majorly classified under the category of machine learning based modelling. Post the exploitation of the parameterizing Active Regions employing the line of sight magnetograms, (Ahmed *et al.*, 2013) employed a black-box approach for classification of the flares, (Yu *et al.*, 2009) on the contrary used C4.5 decision tree to make the model mathematically interpretable. (Jing *et al.*, 2006) employed logistic regression method to achieve comparatively better classification accuracy and (Al-Ghraibah, Boucheron and McAteer, 2015) employed a relevance vector machine for the objective to predict the existence of the flare.

(Qahwaji and Colak, 2007) revisited the (McIntosh, 1990) old research for the sunspot groups and solar cycle classification and employed a support vector machine as dealing with data of higher dimensionality and Cascading-Correlation Neural Network technique for the process of prediction. (Bobra and Couvidat, 2015) amalgamated the concepts of Active-Region attributes derived from vector magnetograms and used SVM for the prediction of existence of the flares. (Nishizuka *et al.*, 2017) employed both vector magnetograms approach and line of sight approach and then compared the performances of the machine learning classifier such as k-NN, Support Vector Machine and Extremely Randomised Tree.

1.2 Literature on Feature Selection and Feature Engineering

Recently the researchers stated pondering upon the concepts of feature selection and feature engineering in the quest to predict the solar flares (Leka and Barnes, 2003a). LDA was employed for feature selection and was applied to 35 feature vectors which resulted in achieving a prediction error if 0 with combination 8 features resulting in a smaller feature space. The same methodology was used in (Welsch *et al.*, 2009) to extract the features for a number of variations of magnetograms and P.flow fields. In the research no significant difference in climatological skill score was detected in a feature subset larger than three feature vectors.

(Bobra and Couvidat, 2015) For feature selection the fisher criterion was employed on 25 vectors followed by using SVM for prediction of existence of flares. This model achieved its maximum performance by employing only 13 features and using only 4 features did not diminish the performance the model by significant capacity. (Welsch *et al.*, 2009) This research only concentrated on feature selection for LOS magnetograms and LDA. In this research the novelty lies in individual ranking of LOS magnetograms using the concept of

RVM-based prediction approach. (Al-Ghraibah, Boucheron and McAteer, 2015) This research has shown a very sopheticathed method to select the features for better processing of data.

1.3 Research Gaps

In the literature review various research gaps are highlighted and the research gaps are presented as research opportunities in this research proposal:

- First it is essentially noted that no research is done on optimising and fabricating the feature selection method to supplement an increase classification performance
- Second, there is a dire need to investigate a modified feature selection method and an AR model or the autoregressive model like the Non-linear autoregressive neural network for achieving higher prediction accuracy for Solar Flares.
- Third examination of an ARIMA model and LSTM network for understanding the sequence of solar flare spiking and predicting their occurrences and measuring the accuracy of prediction.

2 Research Question

The technologies stationed on the land, space, the astronauts, the aircrew, passengers etc all face the adverse impact of the storm caused by solar radiation. These solar radiations are the buy product of the Solar flares. Hence a precise and accurate forecasting of extra-terrestrial weather conditions is essential to mitigate the adversities caused by the unconducive weather conditions.

All the researches discussed in the literature review are primarily focused on the characterization of the Active regions by using the established techniques like line-of-sight and the vector magnetorams but none of the research focused attention towards the relevance of time series of the Active Regions nor attempted to amalgamate a modified feature selection procedure which works dedicatedly to optimise the classification performance.

Hence the aim of this research is to fill in the gaps and opportunities from the previous existing work which formulates our research questions as follows: -

- *Can a modified feature selection method supplemented with an AR model (the autoregressive model like the Non-linear autoregressive neural network) can achieve a higher prediction accuracy for predicting of next Solar Flare?*
- *Can LSTM network work better than ARIMA model for understanding the sequence of Solar Flare spiking and predicting the occurrences of Solar Flares accurately?*
- *Can a modified feature selection method supplemented with an AR model outperform the performance of black-box approach like CNN and LSTM network?*

3 Proposed Methodology

This research is purely based on discovery of new insights hence Knowledge Discovery in Databases (KDD) approach will be used employed in this research. Knowledge Discovery in Databases is a protocol which follow 5-step procedure (Fayyad and Stolorz, 1997).

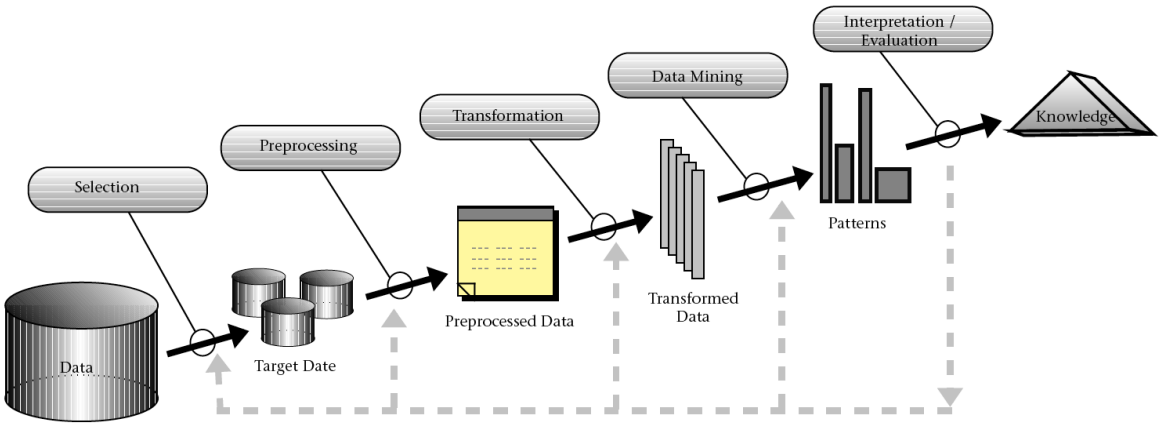


Figure 2: Knowledge Discovery in Databases (KDD)

3.1 Data Source and Dataset

Data for this research will be sourced from multiple sources. The primary source from NOAA SPACE ENVIRONMENT SERVICES CENTER online data repository. Other sources will include Kaggle.com Bigdata Cup Challenge 2019: Flare Prediction an active Big-data Challenge and UCI machine learning repository.

3.2 Data Cleaning and Data Pre-Processing

Steps like removal of special characters from the datasets, dealing with class imbalances by using techniques like ROSE and Smoothing and aggregating the dataset together so that the final dataset is conducive for further feature engineering.

3.3 Feature Extraction from the Dataset

For the black box approaches like LSTM network and Convolution neural network feature extraction is not a necessary step. However other approaches proposed in this research require a methodology from extracting significant feature. A new modified feature extractor will be employed in this research to examine its significance and compare the performance change it induces.

3.4 Data Transformation

During the execution of this research if any hinderance is experienced due to the size or the dimension of the extracted subset then there is a proposal of performing Principal Component Analysis (PCA) which transforms the data set orthogonally without losing the important feature vectors.

4.5 Constructing and Fine Tuning the Proposed Model

First, we will build the baseline model which have been proven to work well historically according to the existing researches. Post building the baseline models according to the literature we will construct the LSTM network or CNN network after successful feature extraction there is a dire need to optimize the hyperparameter for achieving accurate results.

4 Implementation

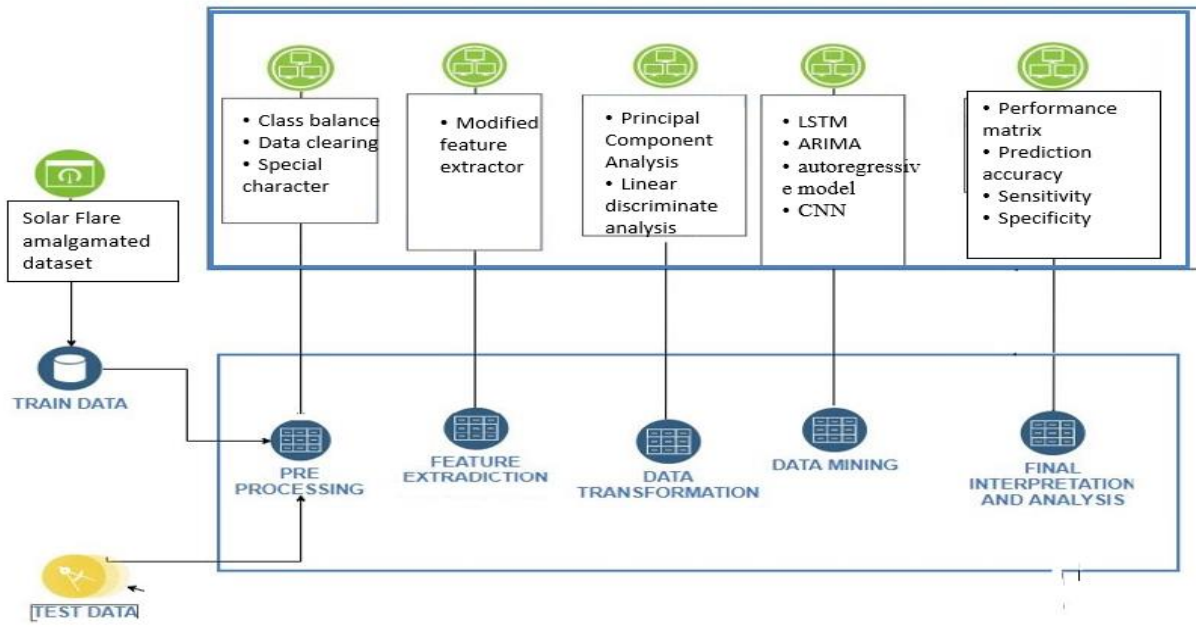


Figure 3: Proposed Implementation

5 Evaluation

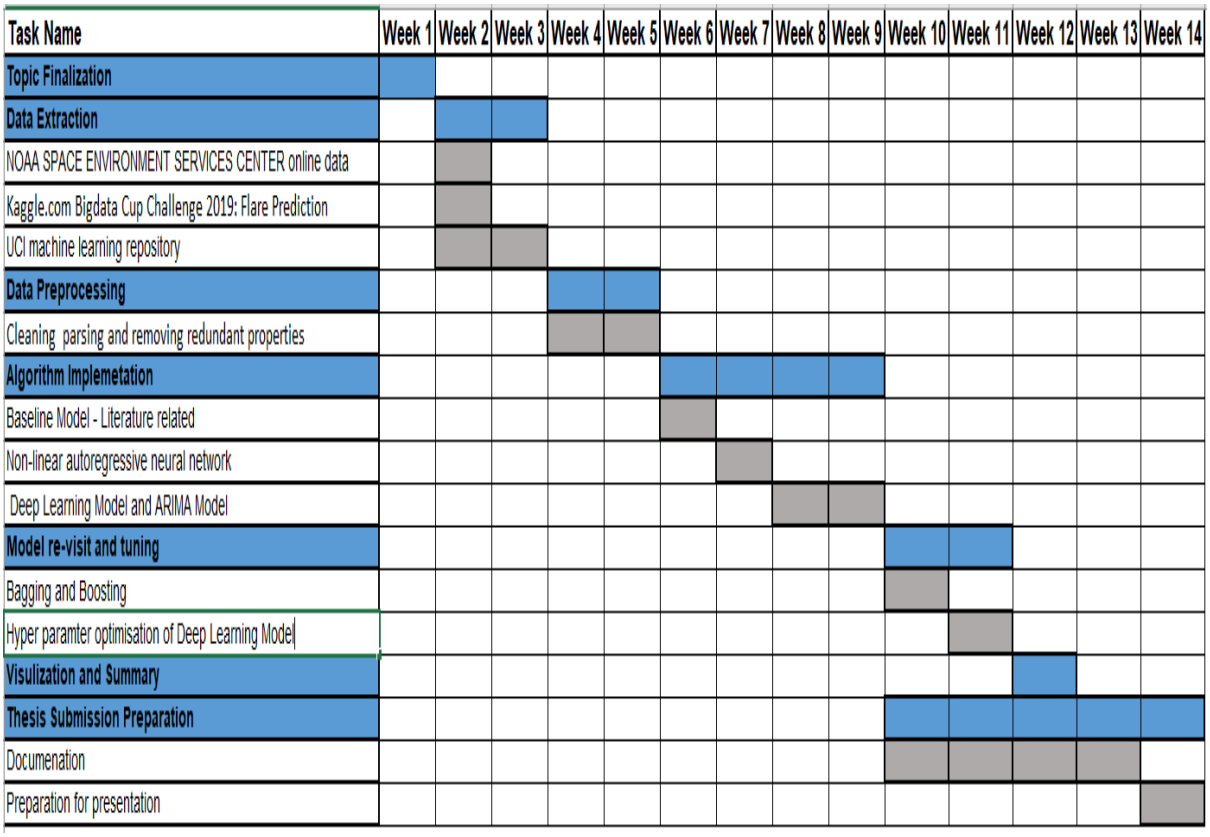
A situation like class imbalance arises while dealing with an aggregated data set, only achieving accuracy is not deterministic of a fair machine learning model and is not a significant metric of a classification model. The elegance of the machine learning model and its robustness will be evaluated by other metrics like Precision, AUC Score, False Negative Rate, F1 measure, False Positive Rate and the accuracy achieved. The main aim to evaluate the robustness of model is to try minimising False Negative Rate and Recall. The proposed model will be evaluated on three benchmarks: -

Benchmark 1: A pure autoregressive model without the key feature extraction. Here it is proposed that autoregressive model without the implementation of precise feature extraction criteria will underperform to predict the existence of Solar Flares.
Benchmark 2: An amalgamation of a modified feature selection method and an AR model or the autoregressive model like the Non-linear autoregressive neural network. Here it is proposed that this amalgamation of modified novel feature extractor along with Non-linear autoregressive neural network will have a significant gain in different performance metrics.
Benchmark 3: Construction and comparison of two diverse model like ARIMA model and LSTM neural network. Here it is proposed that we will evaluate these to diverse model in nature on their ability to understand activity and sequence and timeframe. It will be a fascination experiment to see how a sequence model like LSTM network perform on a time series data nature.
During the process of evaluation, it is essential to take care of concerns like induced bias, cross bias, underfitting and overfitting. A strict train, validation, test data split will be monitor along with the graphical representation so that a regulatory check can be maintained ceasing the chances of any technical induced constraints.

6 Conclusion

Prediction of Solar Flares plays a crucial role to mitigate the adversities caused by the unconducive extra-terrestrial weather conditions. Solar Flares like the X-class are hazardous Solar Flare and the M-class flare Solar Flares which are capable to cause radiation storm and damage on land and in space technologies, hence with the help of data-driven approach we can employ the concepts of machine learning to mitigate the risk induced to help save copious amount of resources and capital (Eastwood *et al.*, 2017). The prediction and detection of Solar Flare is an open challenge in the field of Space Science and Big data analytics. With our novel approach and modelling strategy we aim to fill the gaps in previous researches by amalgamating unorthodox methods for feature selection with Non-linear autoregressive neural network, ARIMA model and deep sequence learning modelling approach LSTM model respectively. In the literature review we have noticed that previous researchers are primarily focused on the characterization of the Active regions by using the established techniques like line-of-sight and the vector magnetograms but none of the research focused attention towards the relevance of time series of the Active Regions and amalgamation of a modified feature selection procedure which works dedicatedly to optimise the classification performance.
Our proposed model will exploit full potential of the time series data to extract numerous new insights which will certainly help to the enhance the performance of the new models and will certainly help to fill the highlighted research gaps.

Appendix 1: GANTT CHART



7 References

Ahmed, O. W. *et al.* (2013) 'Solar Flare Prediction Using Advanced Feature Extraction, Machine Learning, and Feature Selection', *Solar Physics*, 283(1), pp. 157–175. doi: 10.1007/s11207-011-9896-1.

Al-Ghraibah, A., Boucheron, L. E. and McAteer, R. T. J. (2015) 'An automated classification approach to ranking photospheric proxies of magnetic energy build-up', *Astronomy & Astrophysics*, 579, p. A64. doi: 10.1051/0004-6361/201525978.

Alto, P. (1997) 'Selection of Relevant Features'. doi: 10.1016/S0004-3702(97)00063-5.

Bobra, M. G. and Couvidat, S. (2015) 'Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm', *Astrophysical Journal*, 798(2). doi: 10.1088/0004-637X/798/2/135.

Cui, Y. *et al.* (2007) 'Correlation between solar flare productivity and photospheric magnetic field properties II. Magnetic gradient and magnetic shear', *Solar Physics*, 242(1–2), pp. 1–8. doi: 10.1007/s11207-007-0369-5.

Eastwood, J. P. *et al.* (2017) 'The Economic Impact of Space Weather: Where Do We Stand?', *Risk Analysis*, 37(2), pp. 206–218. doi: 10.1111/risa.12765.

Fayyad, U. and Stolorz, P. (1997) 'Data mining and KDD: Promise and challenges', *Future Generation Computer Systems*, 13, pp. 99–115.

Guyon, I. *et al.* (2012) 'Tracking cellulase behaviors', *Biotechnology and Bioengineering*, pp. 1–39. doi: 10.1002/bit.24634.

Hamdi, S. M. *et al.* (2018) 'A time series classification-based approach for solar flare prediction', *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2018-Janua, pp. 2543–2551. doi: 10.1109/BigData.2017.8258213.

Jing, J. *et al.* (2006) 'Erratum: "The Statistical Relationship between the Photospheric Magnetic Parameters and the Flare Productivity of Active Regions" (ApJ, 644, 1273 [2006])', *The Astrophysical Journal*, 652(2), pp. 1796–1796. doi: 10.1086/508989.

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

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

Mason, J. P. and Hoeksema, J. T. (2010) 'Testing automated solar flare forecasting with 13 years of michelson doppler imager magnetograms', *Astrophysical Journal*, 723(1), pp. 634–640. doi: 10.1088/0004-637X/723/1/634.

McIntosh, P. S. (1990) 'The classification of sunspot groups', *Solar Physics*, 125(2), pp. 251–267. doi: 10.1007/BF00158405.

Nishizuka, N. *et al.* (2017) 'Solar Flare Prediction Model with Three Machine-learning Algorithms using Ultraviolet Brightening and Vector Magnetograms', *The Astrophysical Journal*. IOP Publishing, 835(2), p. 156. doi: 10.3847/1538-4357/835/2/156.

Qahwaji, R. and Colak, T. (2007) 'Automatic short-term solar flare prediction using machine learning and sunspot associations', *Solar Physics*, 241(1), pp. 195–211. doi: 10.1007/s11207-006-0272-5.

Townsend, L. W. *et al.* (2006) 'The Carrington event: Possible doses to crews in space from a comparable event', *Advances in Space Research*, 38(2), pp. 226–231. doi: 10.1016/j.asr.2005.01.111.

Welsch, B. T. *et al.* (2009) 'What is the relationship between photospheric flow fields and solar flares?', *Astrophysical Journal*, 705(1), pp. 821–843. doi: 10.1088/0004-637X/705/1/821.

Yu, D. *et al.* (2009) 'Short-term solar flare prediction using a sequential supervised learning method', *Solar Physics*, 255(1), pp. 91–105. doi: 10.1007/s11207-009-9318-9.