Localized Magnetic Substorm Forecasting using Machine Learning

Pascal Sado[1], Lasse Boy Novock Clausen[1], Wojciech Jacek Miloch[1], and Hannes Nickisch[2]

[1]University of Oslo
[2]Philips Research

July 24, 2023

Localised Magnetic Substorm Forecasting using Machine Learning

P. Sado[1], L. B. N. Clausen[1], W. J. Miloch[1], H. Nickisch[2]

[1]Department of Physics, University of Oslo, Oslo, Norway
[2]Philips Research, Hamburg, Germany

Key Points:

- New model combining global forecasting of substorms based on solar wind and local forecasting based on all sky imagers was created
- Previous global substorm forecasting study was successfully reproduced as baseline comparison for new model
- Combined forecasting model performed below necessary precision for scientific purposes

Corresponding author: Pascal Sado, `Pascal.Sado@fys.uio.no`

Abstract

We use a prevailed technique to extract image features and classify 4 seasons of aurora all sky images, combine these with solar wind and interplanetary magnetic field (IMF) data and use this as a basis to forecast the onset of geomagnetic substorms local to the imager. To prove the viability of our model, we successfully reproduce the results of a previous study which used only solar wind and IMF data to forecast global substorm onsets. Although this viability test proves successful and we independently confirm the previous model, our expanded model fails to deliver the necessary performance required for it to be used for accurate localised substorm forecasting.

Plain Language Summary

The solar wind's interaction with the Earth's magnetic field can not only cause beautiful displays of nature, but also create harmful environments for modern infrastructure. Satellite navigation, flights, communication or electric infrastructure can be disrupted or even damaged during strong events. For damage mitigation and research, it is important to be able to forecast the time and location of such occurrences. Our model takes satellite data which has proven to be able to forecast the events globally and supplements these with local imager data to create a localised forecast.

1 Introduction

The solar wind and the interplanetary magnetic field (IMF) are the driving force of space weather around the Earth. Much like regular weather on the Earth, space weather can impact our life. Atmospheric heating and expansion will cause drag on satellites (Marcos et al., 2010), geomagnetically induced currents can disrupt or damage electrical or communication infrastructure (Pirjola, 2000) and ionospheric disturbances will affect the global navigation satellite system (Kintner et al., 2007). Although the effects of space weather storms can be mitigated, they can cause lasting damage. Being able to forecast when extreme space weather events will occur, will not only help with impact mitigation but can also lead to new scientific discoveries, because observations of such events can be planned and targeted.

The aurora is an immediately observable consequence of space weather. When charged particles precipitate onto the Earth, they excite particles in the atmosphere, which in turn release their energy in form of visible light. Different physical processes can cause different auroral morphology, which makes them interesting to study phenomena in the upper atmosphere (Knudsen et al., 2021). Early observations of aurora for study of substorms were performed by Akasofu (1964) and Akasofu et al. (1965) followed by satellite observations later (McPherron et al., 1973). These studies identified the solar wind as the main driving force of substorms (Caan et al., 1975) and developed a model identifying the substorms "growth", "expansion" and "recovery" phases. In this cycle, energy is first stored in the Earth's magnetotail, then suddenly released in the expansion phase before the whole system returns to its resting state.

The main driving force of the growth phase energy storage is to be believed the coupling of the IMF with the Earth's magnetic field, although P. T. Newell and Gjerloev (2011a) and P. Newell et al. (2016) found a strong contribution of the solar wind velocity. Some substorms are reported to have occurred under quiet conditions as well (Russell, 2000b and Miyashita et al., 2011 and Lee et al., 2010). The driving factor for triggering the expansion phase was first believed to be externally through the IMF B_z component (Russell, 2000a) however, recent studies dispute this and found the triggering mechanism to be internally (Freeman & Morley, 2009 and P. T. Newell & Liou, 2011 and Johnson & Wing, 2014).

Visually, a substorms manifests in a specific sequence of morphology in the visible aurora. The aurora progresses from a single east-west arc during quiet times to a brightening and widening band that expands polewards with westward travelling folds before breaking up into smaller and more chaotic structures after which it returns to its quiet state (Akasofu, 1964). This yields an easy way to visually identify the occurrence of substorms as performed by Frey et al. (2004) and Liou (2010). This method can only identify substorms during which visual observations were done . The geomagnetic footprint caused by the substorm allows for automated identification of substorms based on local measurements of the Earth's magnetic field (Forsyth et al., 2015 and P. T. Newell & Gjerloev, 2011a and Ohtani & Gjerloev, 2020) which is a more comprehensive method. The whole field however lacks a single, unified definition and method of identification for substorms.

Based on these methods, lists of substorms were compiled for use in scientific studies. In turn , efforts to forecast substorms have been undertaken. Recently, Maimaiti et al. (2019) have developed a neural network for the binary classification task of whether a substorm will occur anywhere in the Northern Hemisphere's nightside auroral oval within the next hour based on two hours of satellite observations measuring the interplanetary magnetic field and the solar wind. Similarly, Sado et al. (2023) predicted substorm onsets based on images classified using a machine learning algorithm developed by Sado et al. (2022). This method however works locally, based around the location the images have been acquired.

Both methods have their advantages and drawbacks. The first method offers global, almost uninterrupted coverage and offers high precision and recall for forecasting the onset of substorms. It can however not predict the location of occurring events. The second method is trained on images and offers localised forecasting, but is less precise than the global forecasting method.

A method merging the two approaches could inherit both of the advantages of the methods with none of the drawbacks. Being able to precisely forecast the time and location of a substorm would mean that they can be studied better in the future, for example by adjusting cameras, flight paths of satellites or even launch rockets at the correct place and time.

In this work we will attempt to merge these methods to achieve **loc**alised m**a**gnetic subs**t**orm for**e**casting (LOCATE). We will first build a new model that can be trained with data for global forecasting, which we have reproduced independently based on the method by Maimaiti et al. (2019). This data will be fused with local image data and the same training and testing operations will be performed. We then discuss both advantages and limitation of such an approach.

2 Data Sources and Preparation

The data used in this manuscript is threefold. We use satellite data measuring the IMF and solar wind to get global coverage, all sky imager data taking pictures of the aurora from the ground to get local coverage and the SuperMAG list of substorms for our labels.

The IMF and solar wind data are gathered in the OMNI databse (Papitashvili et al., 2014 and Papitashvili & King, 2020). The data are time-shifted to the bowshocknose such that no further processing is necessary. It is provided at 1 min resolution. To avoid small periods of time with missing data, gaps of up to 11 min are filled by linear interpolation. Time series with a length of 120 min will be used later. Interpolating up to 11 min at a time makes up at less than 10% of our data. This way smaller gaps in the data are avoided without sacrificing the integrity of our data as a whole.

All sky imager data are taken from the imager site in Gillam, Manitoba located at N $56° 20.24'$, W $94° 42.36'$. During regular operations, one image is taken every 3 s. Some images may be missing due to data corruption or interrupted coverage. Because the solar wind and IMF data is only available at 1 min resolution, when the image taken closest in time to the satellite data is used it may have been taken up to 30 s earlier or later. The images are preprocessed and their features extracted according to Sado et

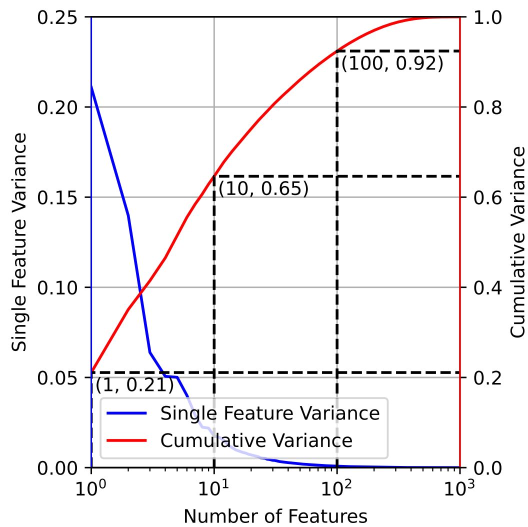


Figure 1: Variance of image features after PCA has been applied. The first 10 features represent 65% of the variance; the first 100 features 92% of the variance in the data.

al. (2022) who have shown that the features extracted by a pretrained neural network for image classification can contain information of physical value. Additionally, principal component analysis (PCA) is employed to reduce the amount of extracted features from 1000 to 10 for the images. As shown in figure 1, this accounts for 65% of the variance in the data. This way some information contained in the data is lost, but the problem commonly referred to as the "Curse of Dimensionality" (Hughes, 1968) is avoided. It means that in order to increase performance of an algorithm such as our classifier, more features can only be added up to a certain point. After this threshold is reached, more data are needed in order to be able to use this information, or degradation of performance is suffered otherwise.

Lastly, we obtain the list of substorms prepared by P. T. Newell and Gjerloev (2011a) based on the SMU and SML indices. These indices are SuperMAG adaptions of the traditionally used auroral electrojet indices. This list is a simple compilation of substorm occurrences including their time of occurrence and location of the magnetometer station where the substorm was identified. See P. T. Newell and Gjerloev (2011a, 2011b) for a detailed explanation of how the list was created. Because we are only interested in substorms in the vicinity of the imager, all substorms that are outside a 10° radius of the imager are discarded. This corresponds to the imager's field of view at a projected altitude of 110 km .

When reproducing the method developed by Maimaiti et al. (2019) we use the same constraints as mentioned in their paper, namely restricting ourselves to substorms occurring in the Northern Hemisphere's nightside auroral oval between 19:00 and 05:00 magnetic local time and between 55° and 75° magnetic latitude. We do not remove outliers for strong SuperMAG electrojet index (SME), since they make up only about 1% of the total data.

2.1 Data Flow and Partitioning

How a piece of data used to train or test the model looks like is shown in figure 2. The upper two panels show IMF and solar wind data, the bottom panel shows the ten most prominent components extracted by PCA stacked on top of each other for easy visualisation. The input matrix for the model consists of these values stacked into a 15x120 matrix (15 variables, for 120 minutes).

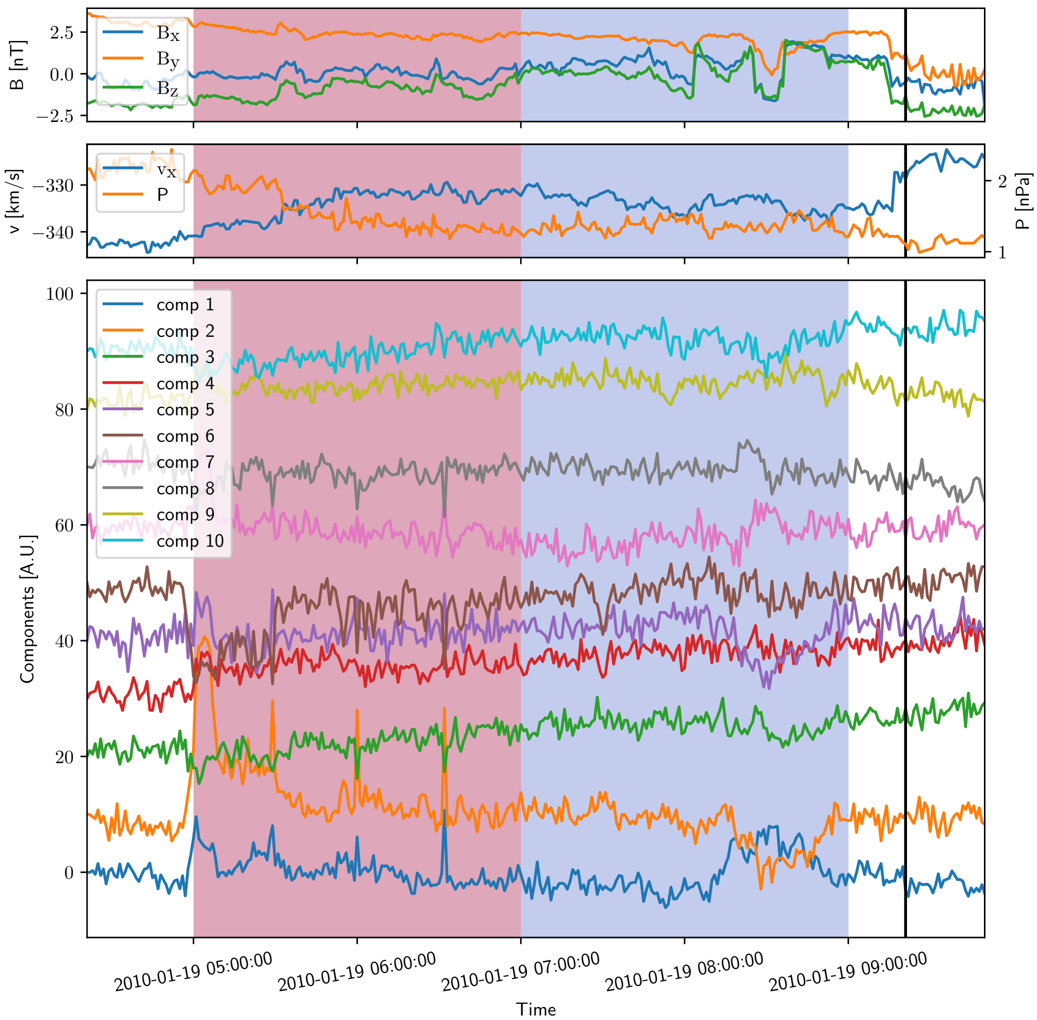


Figure 2: Visualisation of how a sequence of data passed into the neural network looks like. The top panel shows the IMF values, the second panel the solar wind pressure and speed and the last panel the ten most prominent features extracted by PCA. For better illustration they have been offset vertically by a constant value of 10 between each feature. The vertical black line denotes a substorm occurrence. The blue shaded area is followed by a substorm and will be labelled "True", the red shaded area is too far before the substorm and will be labelled "False".

An input interval is labelled "True" if the substorm's occurrence is after the end of the interval and the time between the end of the interval and the occurrence of the substorm is less than or equal to 60 min. An input interval, where the substorm occurs within the interval itself will hence be labelled "false" unless there is another substorm occurring within 60 min afterwards. A substorm occurs at 09:21. The 2 hour long sequence with a blue shadow will be assigned a "True" label because the next substorm occurrence is less than an hour from the end of the sequence, whereas the sequence with the red shadow will be assigned the label "False" because it will not be followed by a substorm.

In figure 3, the flow of data throughout the project is shown. We use the pretrained classifier developed by Sado et al. (2022) to classify the images into the six classes "arc", "diffuse", "discrete", "cloud", "moon" and "clear". Images that are classified to be cloudy

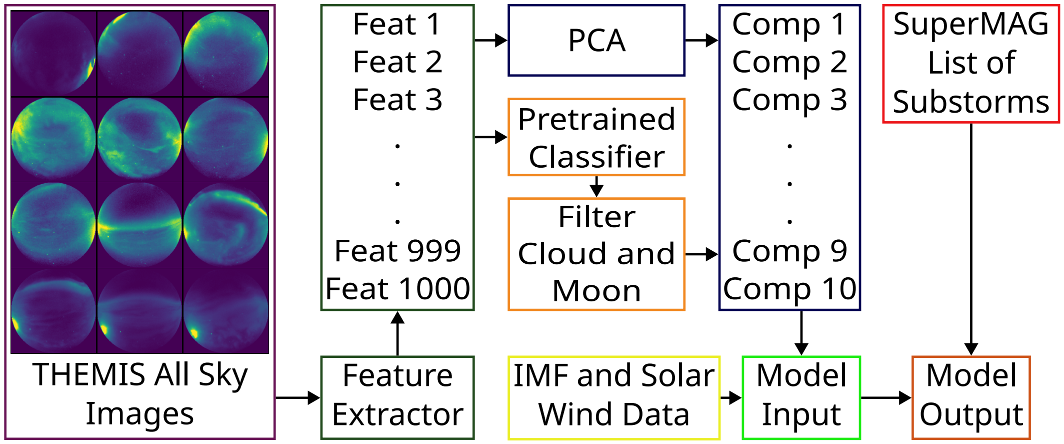


Figure 3: Flow of data in the project. Images are classified using the pretrained classifier developed by Sado et al. (2022). Based on the classifier's output, cloudy and images with the moon visible are discarded. The extracted images' features are reduced to their 10 most prominent components using PCA for better handling and to reduce the dimensionality of the data. IMF and solar wind data are added to 1 min resolution image data. 120 min of data are used to forecast whether a substorm onset will occur within 60 min.

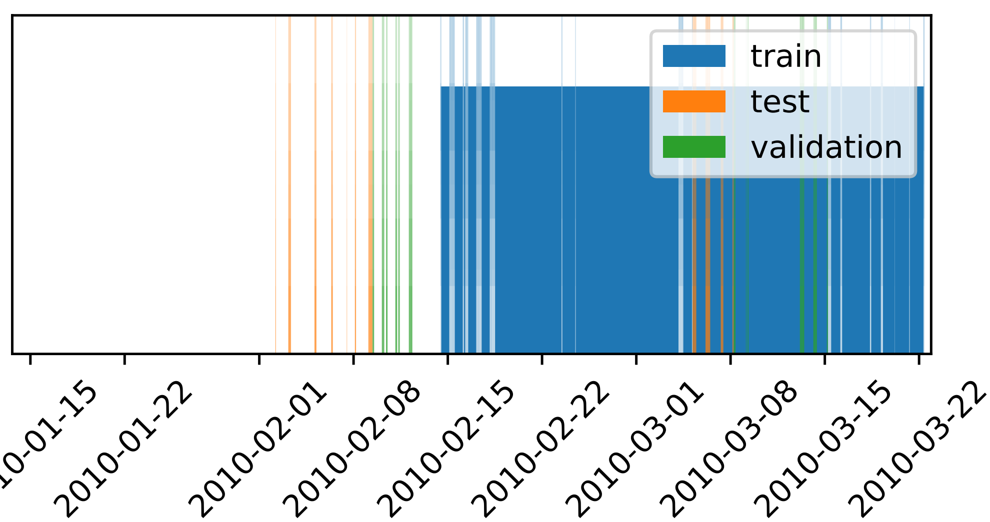


Figure 4: Visualisation of how train, test and validation data are split into several sequential series. The whole set of data is split into 10 sequential chunks, afterwards each chunk is split into 60% train, 20% test and 20% validation data each. This ensures that the distribution of data between train, test and validation is similar and that there is no overlap between the data. Only a part of the data is shown for ease of visualisation, but this principle applies to the whole dataset.

or with the moon visible are removed from the dataset. The moon is too bright to take proper pictures and clouds obscure the aurora, these images therefore contain no information that are useful for forecasting substorms and could lead to unforeseen problems or biases. The numerical features that are extracted on a per-image basis in this process have been shown to be of physical value and can for example be used to model the magnetic footprint of aurora (Sado et al., 2022). We use PCA to reduce the dimensionality of the data and fuse the images' feature data with solar wind and IMF data to build the model's input matrix. The model's output labels are based on the SuperMAG list of substorms (P. T. Newell & Gjerloev, 2011a).

Because so little data are available, we cannot retain a whole season for validation and testing each, instead the data is split into 10 sequential folds, each of which is split sequentially into 60% training and equal amounts of validation and test data. This way, we ensure that seasonality due to the Earth's seasons, the solar cycle and the solar wind (see Lockwood, Mike et al. (2020) and Zhao and Zong (2012)) is equally represented in the training and testing datasets without splitting the data randomly and risk information bleeding from the training into the testing data. This is shown in figure 4. The figure only shows a part of the available data.

Because there is a strong imbalance between negatively labelled ("No Substorm") and positively labelled ("Substorm") of about 20:1 points in the training dataset, the model will tend to value negative results more than positive events. To overcome this problem, the negative cases are randomly undersampled in the training sets, but the validation and test sets are untouched, to properly represent the distribution of substorms as they occur under real-world conditions . The split of training data is used to train the model, validation data will be used for hyperparameter tuning and the test set to evaluate the final model.

3 Model Architecture

There is a significant discrepancy between the amount of data available for the method developed by Maimaiti et al. (2019) for global substorm forecasting and the data available for local substorm forecasting. Our model will have to be smaller to avoid overfitting or bias, but complex enough for the overall task. To ensure that the model we choose for our new task is a generally good model for time series forecasting of substorms, we will first use it for the task of global forecasting.

Deep Residual Networks (ResNet) (He et al., 2016) are a type of convolutional neural network that were first developed to solve image recognition and classification tasks. Their strength lies in their ease of optimization even for deep networks and that they are easy to modify and expand without causing negative side effects. They learn to recognise large scale structures in the first layers and smaller structures in the later layers. The difference between images and time series data as input is not large. Images are of 3 dimensions (width x height x channels), our data has two dimensions (time x features). For an image, different colours represent different features the same way different measurements represent different features in our data. The network's task of classifying based on a time series as compared to an image is therefore relatable. Still, different tasks require different parameters in the design of the network.

ResNets consist of several units with several groups of convolutional layers in each unit. More units or more groups per unit increase the complexity of the network. In order to have a comparable baseline with the original method, we will also use a ResNet that consists of two units, but we decrease the groups per unit to two from three. Instead of developing our own network architecture, we use an architecture developed by Hong et al. (2020) for time series prediction of medical data. Information about the architecture including code to replicate the exact network with trained weights can be found in the code and data we provide alongside the publication.

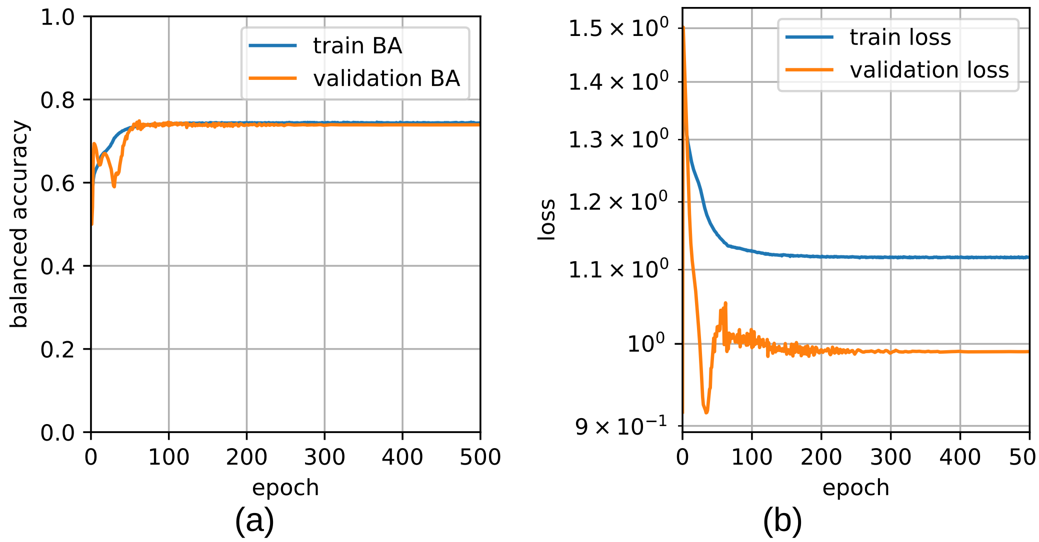


Figure 5: Balanced accuracy (a) and loss (b) during training for the replicated model. Train and validation data are similar and there is no overfitting taking place. The model finishes learning after approximately 300 epochs.

4 Results and Discussion

4.1 Comparison of Models

In table 1 we give an overview of the differences between the used models' architectures, data and results. Our model that reproduces the model developed by Maimaiti et al. (2019) has been kept as close as possible to their model in terms of data, size and capabilities while stile making it possible to integrate the image data into a model of the same architecture.

Some values in the table were not reported in the original publications but could be inferred from the reported results. We will discuss these results in detail below.

4.2 Reproduced Model

Figure 5 shows how the balanced accuracy (5a) and loss (5b) develops during training of the replicated model. The model takes about 300 epochs to settle into a steady state after which no more improvement is taking place. For both training and validation data, the balanced accuracy has settled in at 74%. The balanced accuracy (BA) is calculated like accuracy but each class's contribution is weighted based on the class's occurrence. In a very unbalanced dataset like ours if the model simply classified everything as "False", it would achieve 95% accuracy, but only 50% balanced accuracy. Precision and recall for the positive class are 41% and 63% respectively. Precision is calculated as the true positive cases over all positive predicted cases, i.e. how many of the predicted positive cases are correct, recall is the fraction of positive cases identified of all cases. F1-score is defined as two times the product of precision and recall divided by their sum. This is therefore another metric of accuracy of a model that is based on the two metrics that themselves interest us the most and it is a good metric in general for imbalanced datasets . Because the validation and test sets in Maimaiti et al. (2019) were stratified, we have to calculate balanced precision, recall and F1-score to obtain comparable results. Our model does not perform worse overall than the reproduced model and we therefore confirm the findings of this publication and the viability of the model. However, accounting for real-world conditions by not balancing the test and validation sets,

	Maimaiti substorm onset forecasting	reproduced Maimaiti method	substorm onset forecasting from predicted image classes	combined substorm onset forecasting
feature base	IMF B and solar wind v_x and N_p	IMF B and solar wind v_x and N_p	Predicted image classes aggregated into 5 minute bins	extracted image features, IMF B and solar wind v_x and N_p
feature Length and resolution	120 minutes in 1 minute steps	120 minutes in 1 minute steps	60 minutes in 5 minute steps	120 minutes in 1 minute steps
label base	Substorms identified by Gjerloev (2012) in the nightside auroral zone, outliers removed	substorms identified by Gjerloev (2012) in the nightside auroral zone	substorms occurring within 10 degrees of Gillam ASI identified by Forsyth et al. (2015) and Ohtani and Gjerloev (2020)	substorms occurring within 10 degrees of Gillam ASI identified by Gjerloev (2012) in the nightside auroral zone
label length and resolution	substorm within 60 minutes at every half hour	substorm within 60 minutes at every half hour	substorm within 30 minutes at every 5 minutes	substorm within 60 minutes at every minute
predictor type	1D ResNet	1D ResNet	linear ridge classifier	1D ResNet
amount of parameters	51598	34658	24 + 3 hyperparameters	39618
years of data	all of 1997-2017	all of 2000-2020	aurora seasons 2009/2010, 2010/2011, 2014/2015 and 2015/2016	aurora seasons 2009/2010, 2010/2011, 2014/2015 and 2015/2016

metrics	Maimaiti Validation		Maimaiti Test		reproduced Validation		reproduced Test		image classes Test		combined Validation		combined Test	
label	negative	positive	negative	positive	negative	positive	negative	positive	negative	positive	negative	positive	negative	positive
support	4496	4496	4607	4607	48982	8284	49102	7440	5774	106	12579	481	12577	483
balanced accuracy	0.76		0.74		0.74		0.74		0.66		0.68		0.50	
precision	-	-	-	-	0.93	0.41	0.94	0.35	0.99	0.34	0.98	0.08	0.96	0.04
recall	-	-	-	-	0.85	0.63	0.81	0.66	0.80	0.39	0.70	0.65	0.70	0.29
balanced precision	0.74	0.78	0.74	0.75	0.70	0.80	0.71	0.78	0.58	0.69	0.66	0.67	0.50	0.51
balanced recall	0.80	0.73	0.75	0.73	0.85	0.63	0.81	0.66	0.83	0.39	0.68	0.65	0.72	0.29
F1 score	-	-	-	-	0.89	0.50	0.87	0.46	0.88	0.06	0.67	0.66	0.81	0.06
balanced F1 score	0.77	0.75	0.74	0.74	0.76	0.71	0.76	0.72	0.68	0.50	0.62	0.45	0.59	0.37

Table 1: Summary of metadata and results for the different models used.

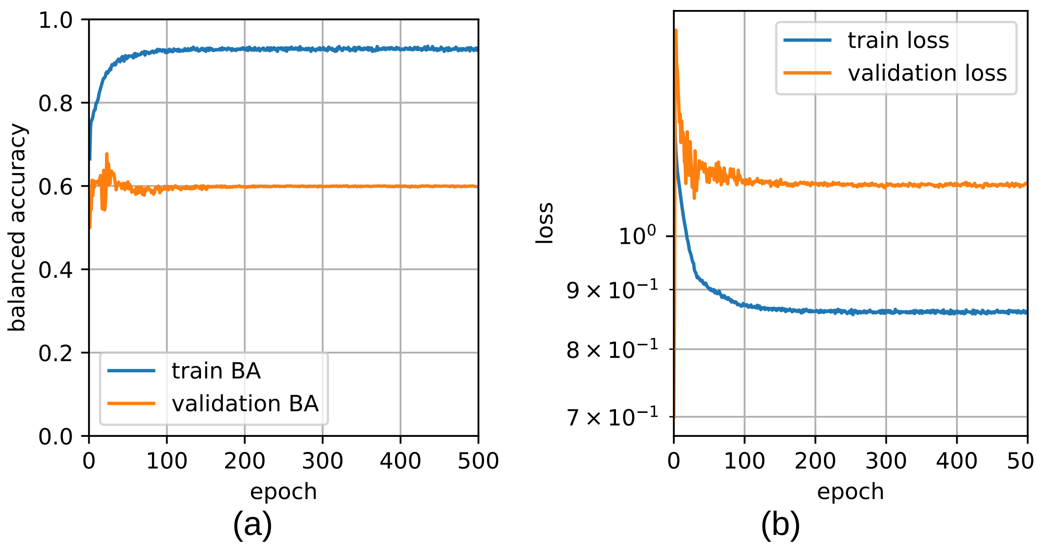


Figure 6: Balanced accuracy (a) and loss (b) during training for the newly created model. Validation data performs worse than the training data. It is difficult to find a configuration where the model generalises and does not overfit. Best performance is achieved at 23 epochs, afterwards it deteriorates and stalls at about 200 epochs.

the precision of the model is worse than previously reported because there are now more false positive cases but the amount of true positive cases stays the same.

4.3 New Model

Figure 6 shows the balanced accuracy (6a) and loss (6b) during training of the newly developed model. The model converges after about 200 epochs for which the balanced accuracy of the validation split achieves approximately 60%. The best result is achieved after 23 epochs with 68% balanced accuracy after which model performance degrades. Figure 7 shows the precision recall curves of the validation (7a) and test (7b) set for the 23rd epoch. Because the dataset is so highly imbalanced, this is a better way to measure the separation of the two classes than a typical ROC curve which plots the true positive rate against the false positive rate. The black line in the figure denotes the relative size of the positive and negative classes at approximately 0.038. If our model was purely guessing, the graph would be equal to this line. As we can see, the validation set exceeds it for higher recall values. This model is chosen as the final model and the test set is evaluated. The model performs barely better than random and only a few events for very low recall values are classified precisely. Comparing this to the results reported by Sado et al. (2023) we see that this model does not outperform a purely imager based forecasting model.

To illustrate how the model underperforms, we have added two keograms with the model's predictions in figure 8. The first (8a) shows an uneventful night on 2009-12-11 where the model falsely predicts an upcoming substorm at approximately 09:00. There is no obvious indication in the data as to why this has happened.

The second selected evening (8b) on 2010-12-31 shows where the model predicted an onset, but fails to precisely identifying the time of the onset. Additionally, this evening illustrates the problem with data procurement as well. Although we already allow for interruptions in the data by interpolating for up to 11 min, there are still moments where data are missing. These small outages cause large gaps in the training and validation data. We performed the same experiment but allowed for more interpolation (up to 30 min) and did not obtain better results.

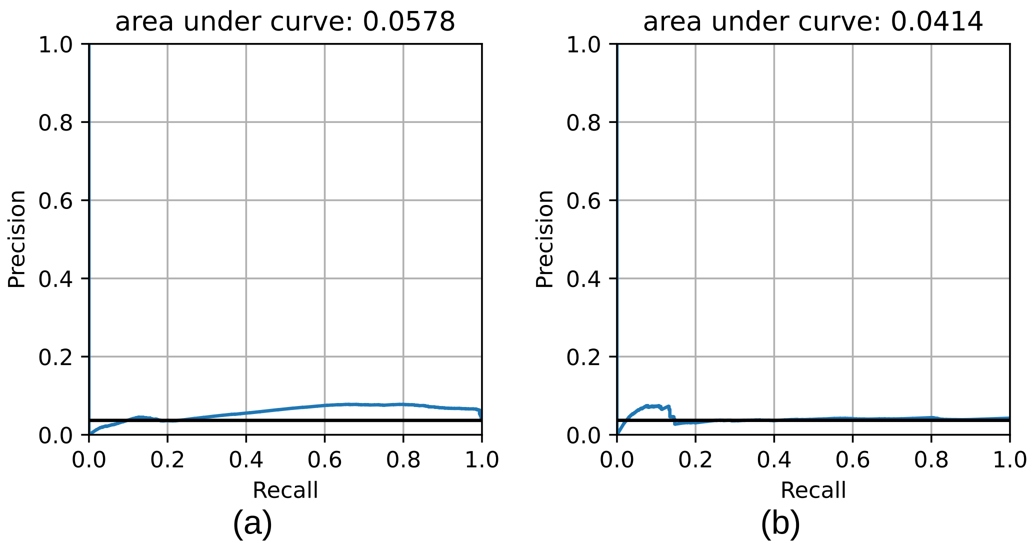


Figure 7: Precision recall curve for validation data (a) and test data (b) for the 23rd epoch of the newly created model. The horizontal black line denotes a model that would be purely guessing. In that case the area under the curve would be 0.038.

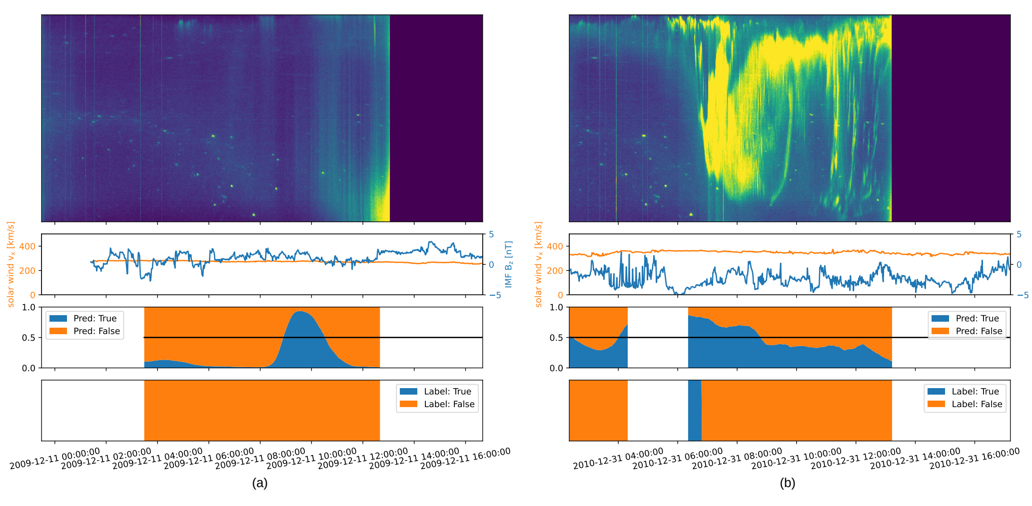


Figure 8: Two keograms with IMF B_z and solar wind v_x plotted underneath. The third panel shows the model's forecasted probability for each time step, the black line denotes the necessary threshold of 50% for the prediction. The bottom panel shows the true label for each point in time. These times were selected for their continuous coverage.

4.4 Failed Attempts

Since we are presenting negative results here, which are still the best of many attempts, we feel obligated to give an overview into the many failed different methods we tried to use:

Oversampling	Simply oversampling the positive class does not yield an improvement.
SMOTE	Synthetic minority over-sampling technique (Chawla et al., 2002) can be used to oversample an underrepresented class in data. Contrary to oversampling, samples are not simply repeated but synthetically created to be similar to known samples but not identical. Both lead to increased overfitting and make it harder for the model to generalise
Different Networks	We create simpler convolutional networks that should be more capable of solving time series forecasting but lack the ability to generalise to other problems however none of them are able to perform to the standards of the model we finally present here. We also try different configurations for the residual units that make up this network.
Class Weights	Different weights for the classes only have the effect that the network is even more likely to classify everything as positive or negative.
PCA	Principal component analysis has a positive effect in that it reduces training time without having a negative impact on the outcome. We believe that when attempting this with more data in the future PCA on the feature space will be an important tool.
IMF interpolation	Increasing the allowed time for IMF interpolation to reduce the amount of outages in the training data increased the amount of available data but does not have a positive effect on the predictive capabilities of the network.
Imager range	Increasing the range of substorms around the imager from $10°$ to $20°$ has no effect.
Hyperparameters	Learning rate and batch size were adjusted by trial and error over several training processes to find the best working combination that allows training without immediate overfitting but still allow the network to learn and generalise.

Overall, we conclude that there needs to be a significant increase in training data for this approach to be feasible.

4.5 Discussion

Our reproduced model confirms the viability of the approach previously demonstrated by Maimaiti et al. (2019). Using deep neural networks is a viable method to forecast the onset of substorms on a global scale and could or should be used in a live environment for space weather forecasts in the future. When reproducing their model, we found that when accounting for more realistic conditions in the validation data, the model's precision is worse than previously reported. The previous model achieved recall rates of 73% at 75% balanced precision, our model obtained 66% recall at 35% precision which increased to 78% when balancing the test dataset. F1-scores were 0.74 for the previous model and 0.46 for our model, increasing to 0.72 when balancing the test set. If a model like this is used in a live forecasting environment it is therefore imperative to remember the limitation in precision of the model and that it will cause many false positive alerts.

In terms of infrastructure, the previous model was written with tensorflow 1.12, ours in pytorch 1.12 and the model consists of about 30% less parameters. This should result in easier deployment and faster training and evaluation times .

A combined approach of using space based solar wind and IMF data together with ground based imager data is not viable to forecast substorms yet. Our findings show that the accuracy of a forecasting model that performs well on just space based data does not translate well onto the combined approach, likely due to the lack of training data which cannot easily be remedied.

To reach the same performance for our local forecasting as was achieved for the global forecasting more data is needed. Most of our data storage-wise comes from processing all sky images.

So far we are using 4 seasons worth of images. Assuming roughly 4 months with 10 h coverage a day out of which half of the images will have to be discarded because of weather, we are left with $4\,\mathrm{season} * 4\,\mathrm{months/season} * 10\,\mathrm{hour/day} * 1/2 = 0.278$ data-years of coverage. Around 72 times as much data, or 288 seasons of all sky imager coverage will be needed to obtain the 20 data-years that were used in satellite data. This would require the processing of roughly $288\,\mathrm{season} * 4\,\mathrm{months/season} * 30\,\mathrm{day/month} * 10\,\mathrm{hour/day} * 60\,\mathrm{images/hour} \approx 21\,\mathrm{M}$ images. Since Themis provides the images online only on a per-hour basis, this would amount to roughly 100 TB of data after download, extraction and storage. Processing is therefore only feasible with direct access to all the data or a combined effort in the space physics community would be required to make the images across different sources available under the same standards. This could for example be realised through a collaborative website where images will be queried by time or predicted image classes. Agreeing on a common feature extractor for prediction would also enable the search for similar images by querying feature space directly. Sharing image features instead of raw image data also serves as a form of data-compression by a factor of ≈ 300.

We still believe that such an approach could yield an improvement to the purely global approach and give a more precise result in terms of time and location for the substorm.

5 Conclusion & Outlook

We combined two methods for the forecasting of substorm onsets, one of which uses IMF and solar wind data to forecast substorms globally and one which uses image data to forecast substorms locally. To show the general capabilities of our combined model, we successfully reproduce the results of the study performing global forecasting and give a better estimate of the model's performance under real-world conditions. Compared to the local forecasting our model performs better but overall it does not manage to reach the necessary performance for it to be deployed in a research environment in a useful manner.

This failure is not with the here-employed method, but rather a lack of training data and the inherent complexity of the problem, which might not be suitable to be described with the used model at the moment. The amount of data would have to be increased about 72-fold and it is therefore not feasible to perform this in this study. Because our model and data are freely and openly available, anyone with access to more or better training data might be able to use this in the future.

Open Research

The data and code for this project are provided on `https://doi.org/10.11582/2023.00023` and `http://tid.uio.no/plasma/LOCATE/` respectively. Both are available under open source licenses.

Acknowledgements

This work is funded in part by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (ERC CoG grant agreement No 866357). The All Sky Image Classifier was made available by Sado et al. (2022) on `http://tid.uio.no/TAME`. We acknowledge NASA contract NAS5-02099 and V. Angelopoulos for use of data from the THEMIS Mission. Specifically: S. Mende and E. Donovan for use of the ASI data, the CSA for logistical support in fielding and data retrieval from the GBO stations, and NSF for support of GIMNAST through grant AGS-1004736. We acknowledge the substorm timing list identified by the Newell and Gjerloev technique (P. T. Newell & Gjerloev, 2011a), the SMU and SML indices (P. T. Newell & Gjerloev, 2011a); and the SuperMAG collaboration (Gjerloev, 2012). The Omni data has been made available on `https://omniweb.gsfc.nasa.gov/ow_min.html` for which we acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service, and OMNI data.

References

Akasofu, S.-I. (1964, April). The development of the auroral substorm. *Planetary and Space Science*, *12*(4), 273–282. Retrieved 2022-07-28, from `https://linkinghub.elsevier.com/retrieve/pii/0032063364901515` doi: 10.1016/0032-0633(64)90151-5

Akasofu, S.-I., Chapman, S., & Meng, C.-I. (1965, November). The polar electrojet. *Journal of Atmospheric and Terrestrial Physics*, *27*(11-12), 1275–1305. Retrieved 2022-07-28, from `https://linkinghub.elsevier.com/retrieve/pii/0021916965900875` doi: 10.1016/0021-9169(65)90087-5

Caan, M. N., McPherron, R. L., & Russell, C. T. (1975). Substorm and interplanetary magnetic field effects on the geomagnetic tail lobes. *Journal of Geophysical Research*, *80*(1), 191–194.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Forsyth, C., Rae, I. J., Coxon, J. C., Freeman, M. P., Jackman, C. M., Gjerloev, J., & Fazakerley, A. N. (2015). A new technique for determining substorm onsets and phases from indices of the electrojet (sophie). *Journal of Geophysical Research: Space Physics*, *120*(12), 10,592-10,606. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JA021343` doi: https://doi.org/10.1002/2015JA021343

Freeman, M. P., & Morley, S. K. (2009). No evidence for externally triggered substorms based on superposed epoch analysis of imf bz. *Geophysical Research Letters*, *36*(21). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009GL040621` doi: https://doi.org/10.1029/2009GL040621

Frey, H., Mende, S., Angelopoulos, V., & Donovan, E. (2004). Substorm onset observations by image-fuv. *Journal of Geophysical Research: Space Physics*, *109*(A10).

Gjerloev, J. W. (2012). The supermag data processing technique. *Journal of Geophysical Research: Space Physics*, *117*(A9). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JA017683` doi: https://doi.org/10.1029/2012JA017683

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778). doi: 10.1109/CVPR.2016.90

Hong, S., Xu, Y., Khare, A., Priambada, S., Maher, K., Aljiffry, A., … Tumanov, A. (2020). Holmes: Health online model ensemble serving for deep learning models in intensive care units. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1614–1624).

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, *14*(1), 55-63. doi: 10.1109/TIT.1968.1054102

Johnson, J. R., & Wing, S. (2014). External versus internal triggering of substorms: An information-theoretical approach. *Geophysical Research Letters*, *41*(16), 5748-5754. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL060928 doi: https://doi.org/10.1002/2014GL060928

Kintner, P. M., Ledvina, B. M., & de Paula, E. R. (2007, September). GPS and ionospheric scintillations: GPS AND IONOSPHERIC SCINTILLATIONS. *Space Weather*, *5*(9), n/a–n/a. Retrieved 2022-08-22, from http://doi.wiley.com/10.1029/2006SW000260 doi: 10.1029/2006SW000260

Knudsen, D. J., Borovsky, J. E., Karlsson, T., Kataoka, R., & Partamies, N. (2021). Topical collection on auroral physics. *Space Science Reviews*, *217*(1), 19.

Lee, D.-Y., Choi, K.-C., Ohtani, S., Lee, J. H., Kim, K. C., Park, K. S., & Kim, K.-H. (2010, January). Can intense substorms occur under northward IMF conditions?: SUBSTORMS UNDER NORTHWARD IMF. *Journal of Geophysical Research: Space Physics*, *115*(A1), n/a–n/a. Retrieved 2023-04-11, from http://doi.wiley.com/10.1029/2009JA014480 doi: 10.1029/2009JA014480

Liou, K. (2010). Polar ultraviolet imager observation of auroral breakup. *Journal of Geophysical Research: Space Physics*, *115*(A12).

Lockwood, Mike, Owens, Mathew J., Barnard, Luke A., Haines, Carl, Scott, Chris J., McWilliams, Kathryn A., & Coxon, John C. (2020). Semi-annual, annual and universal time variations in the magnetosphere and in geomagnetic activity: 1. geomagnetic data. *J. Space Weather Space Clim.*, *10*, 23. Retrieved from https://doi.org/10.1051/swsc/2020023 doi: 10.1051/swsc/2020023

Maimaiti, M., Kunduri, B., Ruohoniemi, J., Baker, J., & House, L. L. (2019). A deep learning-based approach to forecast the onset of magnetic substorms. *Space Weather*, *17*(11), 1534–1552.

Marcos, F., Lai, S., Huang, C., Lin, C., Retterer, J., Delay, S., & Sutton, E. (2010, August). Towards Next Level Satellite Drag Modeling. In *AIAA Atmospheric and Space Environments Conference.* Toronto, Ontario, Canada: American Institute of Aeronautics and Astronautics. Retrieved 2022-08-22, from https://arc.aiaa.org/doi/10.2514/6.2010-7840 doi: 10.2514/6.2010-7840

McPherron, R. L., Russell, C. T., & Aubry, M. P. (1973, June). Satellite studies of magnetospheric substorms on August 15, 1968: 9. Phenomenological model for substorms. *Journal of Geophysical Research*, *78*(16), 3131–3149. Retrieved 2022-07-28, from http://doi.wiley.com/10.1029/JA078i016p03131 doi: 10.1029/JA078i016p03131

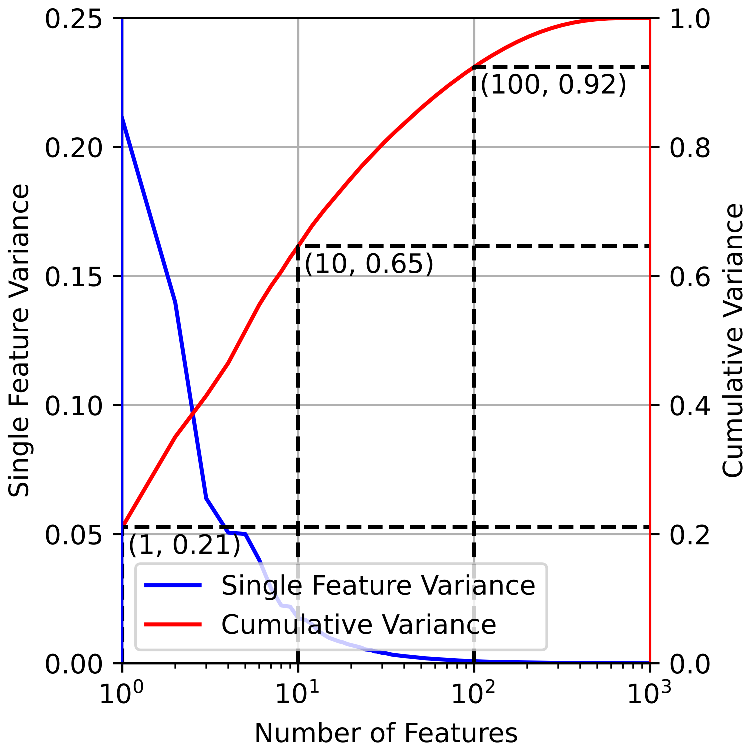
Miyashita, Y., Kamide, Y., Liou, K., Wu, C.-C., Ieda, A., Nishitani, N., ... Mukai, T. (2011, September). Successive substorm expansions during a period of prolonged northward interplanetary magnetic field: SUBSTORMS DURING PROLONGED NORTHWARD IMF. *Journal of Geophysical Research: Space Physics*, *116*(A9), n/a–n/a. Retrieved 2023-04-11, from http://doi.wiley.com/10.1029/2011JA016719 doi: 10.1029/2011JA016719

Newell, P., Liou, K., Gjerloev, J., Sotirelis, T., Wing, S., & Mitchell, E. (2016). Substorm probabilities are best predicted from solar wind speed. *Journal of Atmospheric and Solar-Terrestrial Physics*, *146*, 28–37.

Newell, P. T., & Gjerloev, J. W. (2011a). Evaluation of supermag auroral electrojet indices as indicators of substorms and auroral power. *Journal of Geophysical Research: Space Physics*, *116*(A12). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JA016779 doi: https://

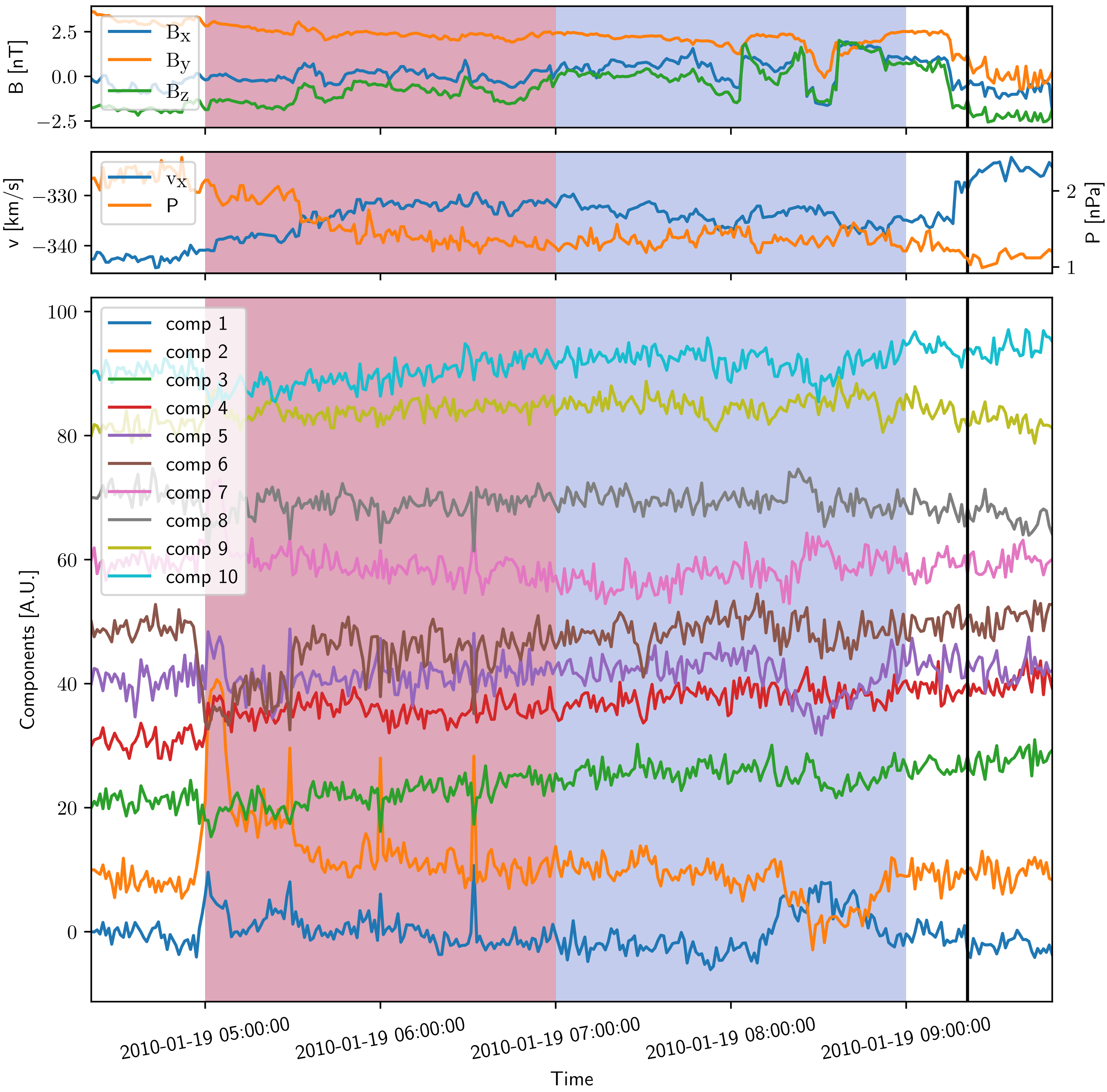
doi.org/10.1029/2011JA016779
Newell, P. T., & Gjerloev, J. W. (2011b). Substorm and magnetosphere characteristic scales inferred from the supermag auroral electrojet indices. *Journal of Geophysical Research: Space Physics*, *116*(A12). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JA016936 doi: https://doi.org/10.1029/2011JA016936
Newell, P. T., & Liou, K. (2011). Solar wind driving and substorm triggering. *Journal of Geophysical Research: Space Physics*, *116*(A3). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JA016139 doi: https://doi.org/10.1029/2010JA016139
Ohtani, S., & Gjerloev, J. W. (2020). Is the substorm current wedge an ensemble of wedgelets?: Revisit to midlatitude positive bays. *Journal of Geophysical Research: Space Physics*, *125*(9), e2020JA027902. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JA027902 (e2020JA027902 2020JA027902) doi: https://doi.org/10.1029/2020JA027902
Papitashvili, N., Bilitza, D., & King, J. (2014, January). OMNI: A Description of Near-Earth Solar wind Environment. In *40th cospar scientific assembly* (Vol. 40, p. C0.1-12-14).
Papitashvili, N. E., & King, J. H. (2020). *Omni 1-min data set [data set].* Accessed on 2023-April-17. Retrieved from https://doi.org/10.48322/45bb-8792 doi: 10.48322/45bb-8792
Pirjola, R. (2000, December). Geomagnetically induced currents during magnetic storms. *IEEE Transactions on Plasma Science*, *28*(6), 1867–1873. Retrieved 2022-08-22, from http://ieeexplore.ieee.org/document/902215/ doi: 10.1109/27.902215
Russell, C. T. (2000a). How northward turnings of the imf can lead to substorm expansion onsets. *Geophysical Research Letters*, *27*(20), 3257-3259. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000GL011910 doi: https://doi.org/10.1029/2000GL011910
Russell, C. T. (2000b, October). How northward turnings of the IMF can lead to substorm expansion onsets. *Geophysical Research Letters*, *27*(20), 3257–3259. Retrieved 2023-04-11, from http://doi.wiley.com/10.1029/2000GL011910 doi: 10.1029/2000GL011910
Sado, P., Clausen, L. B. N., Miloch, W. J., & Nickisch, H. (2022). Transfer learning aurora image classification and magnetic disturbance evaluation. *Journal of Geophysical Research: Space Physics*, *127*(1), e2021JA029683. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JA029683 (e2021JA029683 2021JA029683) doi: https://doi.org/10.1029/2021JA029683
Sado, P., Clausen, L. B. N., Miloch, W. J., & Nickisch, H. (2023). Substorm onset prediction using machine learning classified auroral images. *Space Weather*, *21*(2), e2022SW003300.
Zhao, H., & Zong, Q.-G. (2012). Seasonal and diurnal variation of geomagnetic activity: Russell-mcpherron effect during different imf polarity and/or extreme solar wind conditions. *Journal of Geophysical Research: Space Physics*, *117*(A11). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JA017845 doi: https://doi.org/10.1029/2012JA017845

Figure 1.



0.25
0.20
0.15
0.10
0.05
0.00
1.0
0.8
0.6
0.4
0.2
0.0
10^0
10^1
10^2
10^3
Single Feature Variance
Cumulative Variance
Number of Features
(100, 0.92)
(10, 0.65)
(1, 0.21)
Single Feature Variance
Cumulative Variance

Figure 2.



B [nT]
B_x
B_y
B_z
2.5
0.0
−2.5
v [km/s]
P [nPa]
v_x
P
−330
−340
2
1
Components [A.U.]
comp 1
comp 2
comp 3
comp 4
comp 5
comp 6
comp 7
comp 8
comp 9
comp 10
100
80
60
40
20
0
2010-01-19 05:00:00
2010-01-19 06:00:00
2010-01-19 07:00:00
2010-01-19 08:00:00
2010-01-19 09:00:00
Time

Figure 3.

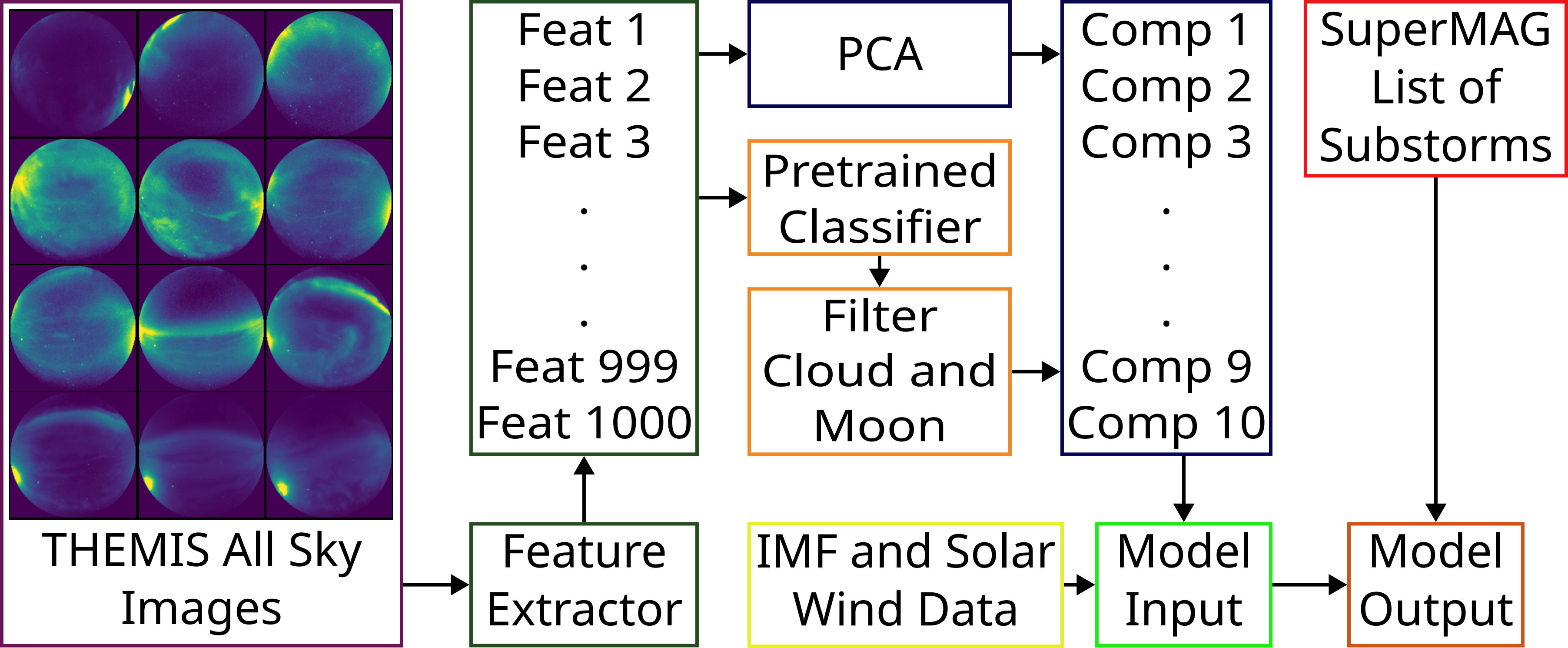
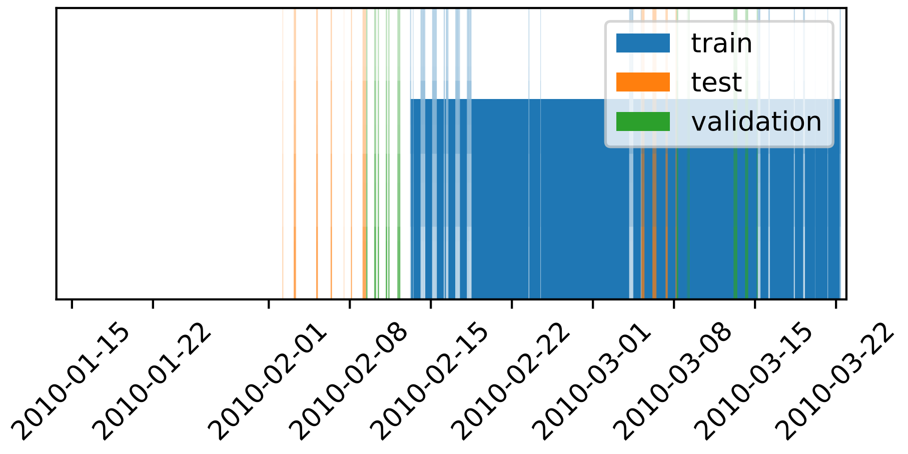
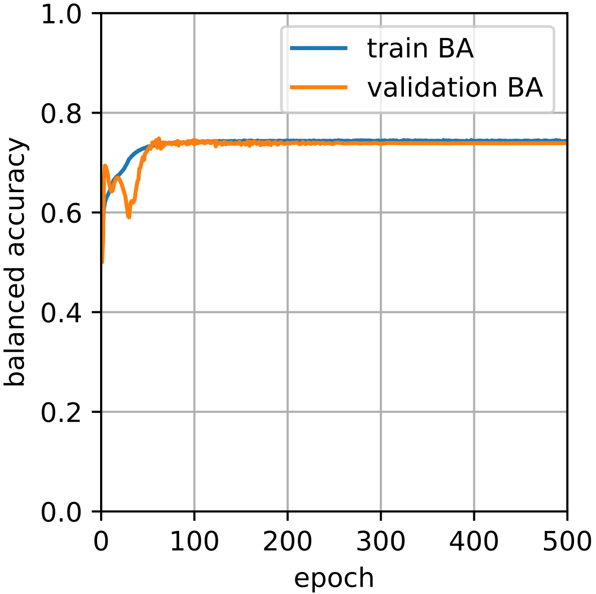

THEMIS All Sky Images
Feature Extractor
Feat 1
Feat 2
Feat 3
.
.
.
Feat 999
Feat 1000
PCA
Pretrained Classifier
Filter Cloud and Moon
Comp 1
Comp 2
Comp 3
.
.
.
Comp 9
Comp 10
SuperMAG List of Substorms
IMF and Solar Wind Data
Model Input
Model Output

Figure 4.

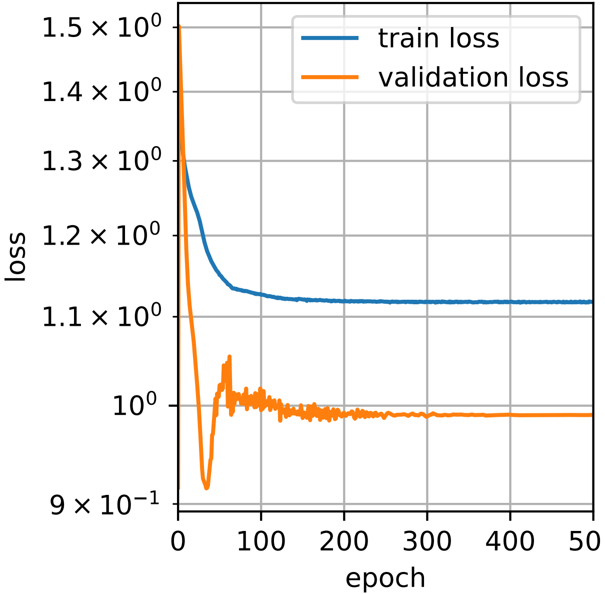


train
test
validation
2010-01-15
2010-01-22
2010-02-01
2010-02-08
2010-02-15
2010-02-22
2010-03-01
2010-03-08
2010-03-15
2010-03-22

Figure 5.

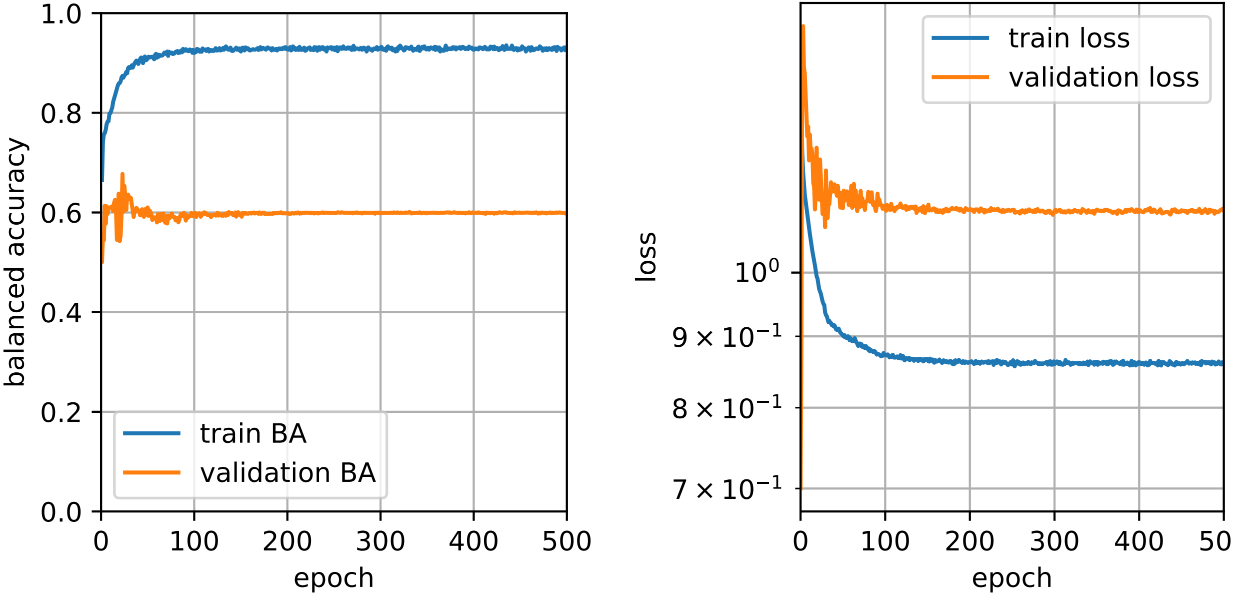

train BA
validation BA
balanced accuracy
epoch
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

(a)


train loss
validation loss
loss
epoch
1.5×10^0
1.4×10^0
1.3×10^0
1.2×10^0
1.1×10^0
10^0
9×10^{-1}
0
100
200
300
400
50

(b)

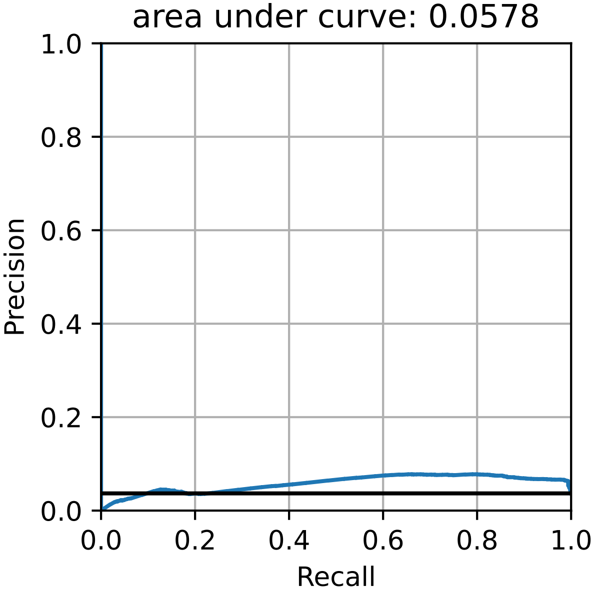
Figure 6.



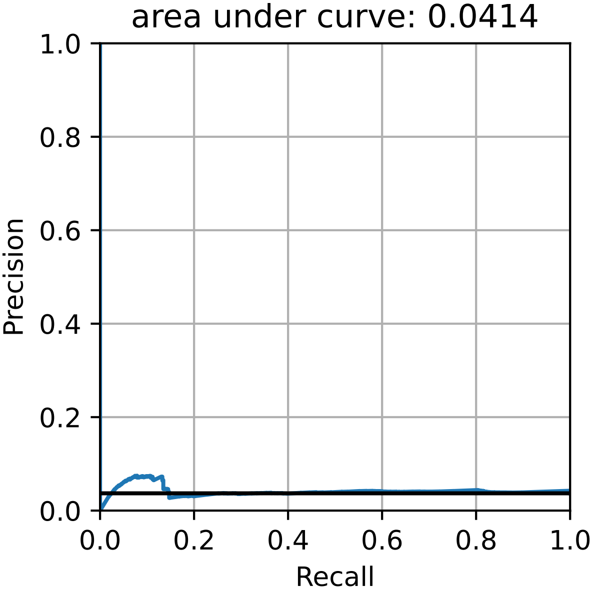
balanced accuracy
epoch
train BA
validation BA
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
loss
epoch
train loss
validation loss
10^0
9×10^{-1}
8×10^{-1}
7×10^{-1}
0
100
200
300
400
50

(a) (b)

Figure 7.

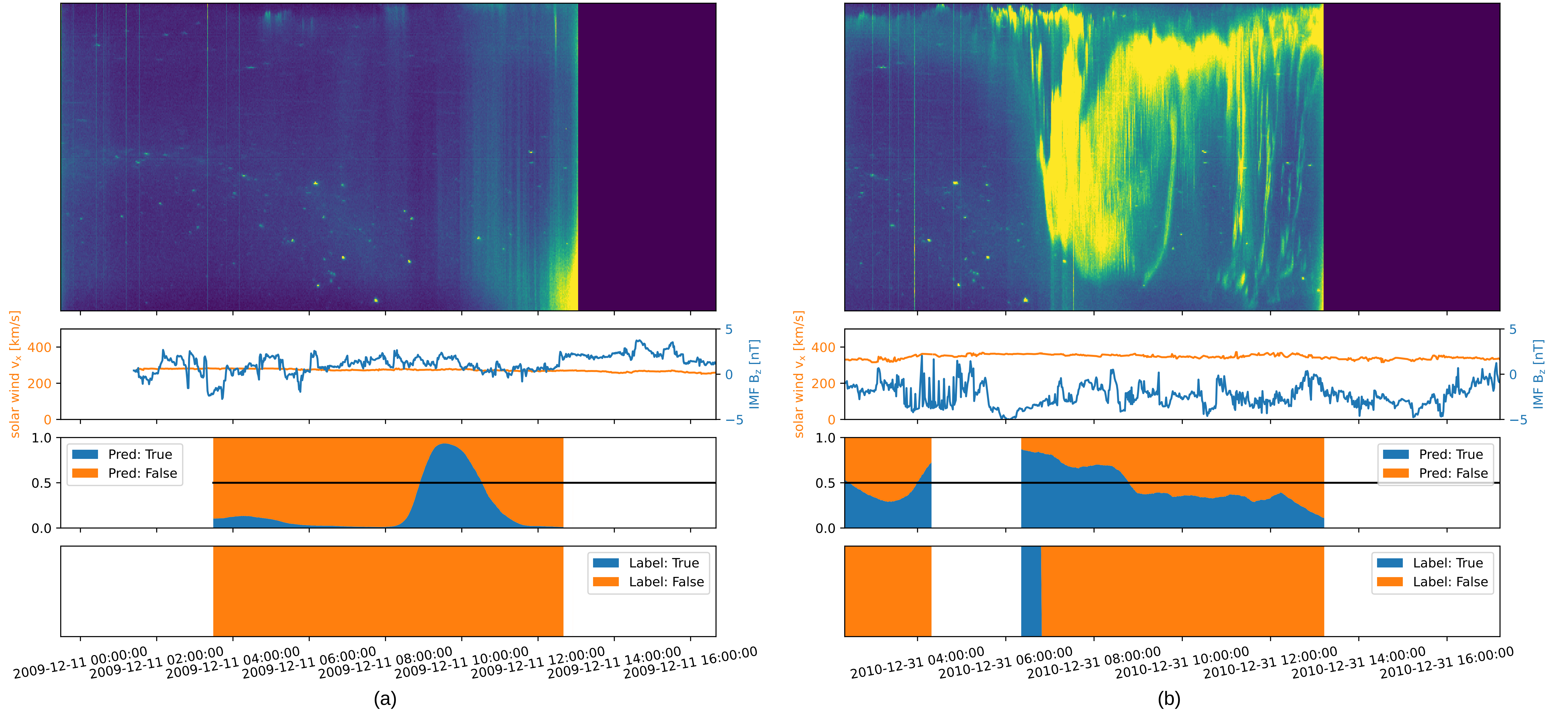

area under curve: 0.0578
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0

(a)


area under curve: 0.0414
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0

(b)

Figure 8.



solar wind v_x [km/s]
IMF B_z [nT]
400
200
0
5
0
−5
1.0
0.5
0.0
Pred: True
Pred: False
Label: True
Label: False
2009-12-11 00:00:00
2009-12-11 02:00:00
2009-12-11 04:00:00
2009-12-11 06:00:00
2009-12-11 08:00:00
2009-12-11 10:00:00
2009-12-11 12:00:00
2009-12-11 14:00:00
2009-12-11 16:00:00
(a)
2010-12-31 04:00:00
2010-12-31 06:00:00
2010-12-31 08:00:00
2010-12-31 10:00:00
2010-12-31 12:00:00
2010-12-31 14:00:00
2010-12-31 16:00:00
(b)