Key Points:
- We present the first deep learning-based time series model to forecast the onset of a magnetic substorm over the next 1 hr
- The model has been developed using a comprehensive list of onsets compiled between 1997 and 2017 and achieves 72% precision and 77% recall
- A significant overlap in solar wind and IMF conditions prior to minor substorms and non-onsets reduces the classification accuracy

Correspondence to:
M. Maimaiti,
rafiq@vt.edu

A Deep Learning-Based Approach to Forecast the Onset of Magnetic Substorms

M. Maimaiti[1] [iD]**, B. Kunduri[1]** [iD]**, J. M. Ruohoniemi[1]** [iD]**, J. B. H. Baker[1]** [iD]**, and Leanna L. House[2]**

[1]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, [2]Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Abstract The auroral substorm has been extensively studied over the last six decades. However, our understanding of its driving mechanisms is still limited and so is our ability to accurately forecast its onset. In this study, we present the first deep learning-based approach to predict the onset of a magnetic substorm, defined as the signature of the auroral electrojets in ground magnetometer measurements. Specifically, we use a time history of solar wind speed (V_x), proton number density, and interplanetary magnetic field (IMF) components as inputs to forecast the occurrence probability of an onset over the next 1 hr. The model has been trained and tested on a data set derived from the SuperMAG list of magnetic substorm onsets and can correctly identify substorms ∼75% of the time. In contrast, an earlier prediction algorithm correctly identifies ∼21% of the substorms in the same data set. Our model's ability to forecast substorm onsets based on solar wind and IMF inputs prior to the actual onset time, and the trend observed in IMF B_z prior to onset together suggest that a majority of the substorms may not be externally triggered by northward turnings of IMF. Furthermore, we find that IMF B_z and V_x have the most significant influence on model performance. Finally, principal component analysis shows a significant degree of overlap in the solar wind and IMF parameters prior to both substorm and nonsubstorm intervals, suggesting that solar wind and IMF alone may not be sufficient to forecast all substorms, and preconditioning of the magnetotail may be an important factor.

1. Introduction

A substorm, often referred to as the magnetospheric substorm or an auroral substorm, is a complex phenomenon involving energy transfer from the magnetotail to the auroral ionosphere. Strong coupling between the solar wind, magnetosphere, and the ionosphere is an important aspect of the substorm phenomenon and its signatures can be observed in several different instruments ranging from satellites in the magnetotail to ground-based all-sky imagers and magnetometers. The duration of a substorm is about 1–2 hr (Akasofu, 1964; Akasofu et al., 1965; McPherron et al., 1973) and is typically categorized into three phases: a growth phase in which energy is extracted from the solar wind and stored in the nightside tail of the magnetosphere; an expansion phase in which stored energy is explosively released; and a recovery phase in which the magnetosphere relaxes to a quiet state (McPherron, 1979).

In all-sky imaging data, a substorm appears as a sudden explosive brightening followed by a poleward expansion of the aurora on the nightside (Akasofu, 1964). This expansion is often associated with a westward ionospheric current flowing across the bulge of the expanding aurora. This westward current forms the ionospheric component of a 3-D current system called the Substorm Current Wedge (SCW), flowing in from the magnetotail on the dawnside and out on the duskside (McPherron & Chu, 2016; McPherron et al., 1973). The westward ionospheric component of the SCW leads to an increase in the westward auroral electrojet, which is measured by the ground magnetometers and is usually observed as a depression in the auroral electrojet indices, such as AL (Auroral Lower; Davis & Sugiura, 1966) and SML (SuperMAG version of AL index; Newell & Gjerloev, 2011a). In this study, we define these signatures of the auroral substorm in ground magnetometers and auroral indices as the magnetic substorm.

Substorms have been extensively studied over the last several decades (Akasofu, 1964; Lui, 2000; Lyons, 1995; McPherron, 1979; Newell & Gjerloev, 2011b; Newell et al., 2016; Pudovkin, 2013) using observations from both space and ground-based instruments, as well as through first principles-based models. In particular, several studies have focused on analyzing the impact of different solar wind and interplanetary magnetic field (IMF) parameters on the substorm growth phase (Akasofu, 1975; Caan et al., 1977; Freeman & Morley,

2004; Henderson et al., 1996; Lyons, 1995; Newell et al., 2016). A majority of these studies find IMF B_z to be the most important factor controlling substorm activity. Specifically, a southward directed B_z was considered to be crucial to enable energy transfer between the IMF and the magnetosphere (Caan et al., 1977; McPherron et al., 1973; McPherron, 1979; Pudovkin, 2013). More recently, Newell et al. (2016) used several years of substorm onsets compiled by SuperMAG (Newell & Gjerloev, 2011a) and showed that solar wind velocity (V_x) could be the most important factor in driving substorms. While the growth phase of substorms was typically associated with periods of elevated solar wind driving, a few studies reported observations of substorms occurring during relatively quiet conditions (Kullen & Karlsson, 2004; Lee et al., 2010; Miyashita et al., 2011). The growth phase of such substorms was thought to be driven by IMF B_y reconnection or a previous southward B_z interval. In summary, while substorms are generally thought to occur during periods dominated by strong geomagnetic acitivity, a subset has been reported to occur even during quiet conditions.

A widely debated topic has been the mechanism that triggers substorm expansion phase onset. Some studies proposed that the onset may be externally triggered by a change in solar wind and IMF conditions, such as a northward turning of B_z (Lyons, 1995; Lyons et al., 1997; Russell, 2000), whereas recent observations contradict these findings and suggest that substorms may be internally triggered and a northward turning of B_z is not necessary (Morley & Freeman, 2007; Johnson & Wing, 2014; Freeman & Morley, 2009; Johnson & Wing, 2014; Wild et al., 2009). Furthermore, understanding the chronological sequence of events constituting the actual trigger process has been a topic of interest for several researchers. Specifically, opposing views have emerged on whether current disruption in the near-Earth region is followed by magnetic reconnection occurring farther out in the magnetotail or vice versa (Baker et al., 1996; Lui, 1991; 1996). The Time History of Events and Macroscale Interactions during Substorms mission (Angelopoulos, 2008) addressed this question by demonstrating that tail reconnection happens prior to current disruption (Angelopoulos et al., 2008). The study also raised important questions about the role of preconditioning of the magnetosphere in terms of tail reconnection during the growth phase.

Over the last few decades, our reliance on satellites for communications, navigation, positioning, meteorology, Earth observation, science, technology, security, and defense has increased significantly. Currently, there are several hundred satellites in geosynchronous orbit and their operations can be perturbed and on occasion totally disrupted by substorms. For example, substorm-injected energetic particles can lead to surface charging in satellites, which can damage surface materials and underlying components (Gubby & Evans, 2002; Horne et al., 2013; O'Brien, 2009). Substorms can also produce geomagnetically induced currents that adversely impact electric supply networks (Belakhovsky et al., 2019; Freeman et al., 2019). Forecasting substorm activity has therefore been a topic of practical as well as theoretical space weather interest. Early efforts utilized neural networks with one to two hidden layers (each layer composed of a few units) to predict the AE index over the next time step, as a proxy for substorm activity, using limited IMF and solar wind data as inputs (Gavrishchaka & Ganguli, 2001; Gleisner & Lundstedt, 1997; Hernandez et al., 1993). More recently, support vector machines were trained on a limited data set (197 data points comprising 98 onsets and 99 non-onset intervals) and achieved 78% accuracy in predicting substorm onsets (Tanaka et al., 2015) (Note that, in contrast to the current study in which we report both precision and recall rates, only accuracy is reported in ; Tanaka et al., 2015). Traditional machine learning algorithms have often been successful in modeling functions between input and output data. However, a drawback associated with those algorithms is the necessity to manually extract features from the data set, a tedious and sometimes impractical task when working with large and complex data sets. In contrast, while deep learning methods require extensive computational resources, they have been successful in modeling complex mappings between input and output variables without the necessity for manual feature engineering, especially in large data sets (LeCun et al., 2015). Finally, a majority of the earlier neural network-based models were trained to predict the auroral indices (Gavrishchaka & Ganguli, 2001; Hernandez et al., 1993) and such models cannot be directly employed to forecast a substorm onset. We believe it would be useful to develop a deep learning model trained specifically for a binary classification task: forecasting whether a substorm will occur over the next hour.

In this study, we present the first deep learning-based approach to predict the onset of a magnetic substorm, as defined by the SuperMAG SML index (Gjerloev, 2012). Specifically, we use a time history of solar wind bulk speed (V_x), proton number density (N_p), and IMF components (B_x, B_y, and B_z) in the geocentric solar

magnetospheric (GSM) coordinate system as inputs to forecast the occurrence probability of an onset over the next 1 hr. The magnetic substorms used for model development comprises a comprehensive SuperMAG list of ~35,000 magnetic onsets compiled between 1997 and 2017. The paper is organized as follows. Section 2 describes the data sets used in this study, gives the definitions of technical terms, and, explains model development and data preprocessing. Section 3 presents model prediction results and section 4 discusses some important findings and relates them to other published works on substorms. Section 5 gives the summary and conclusions of this study.

2. Data Sets and Model Architecture

2.1. Data Sets and Data Reprocessing

Solar Wind and IMF.The solar wind and IMF data were primarily obtained from the Advanced Composition Explorer (ACE) spacecraft in orbit around the Earth-Sun L1 Lagrangian point, which is ~240 Earth radius away from the Earth's center. IMF data are from the magnetic field instrument MAG (Smith et al., 1998) and has 16-s resolution, while the solar wind data are from the Solar Wind Electron, Proton, and Alpha Monitor (SWEPAM) instrument (McComas et al., 1998) and has 64-s resolution. We used 1-min-averaged OMNI SW and IMF observations, which were shifted in time to the bow shock nose see https://omniweb.gsfc.nasa.gov/html/ow_data.html\#time_shift for details.

Magnetic Substorm Onset List and SuperMAG SML Index. We use the comprehensive list of ~35,000 substorm onsets compiled between 1997 and 2017 using the SML index (Gjerloev, 2012) to create our data set. The substorm onsets were identified based on the SML index using the substorm identification criteria proposed by Newell and Gjerloev (2011a). We discard outliers and undesirable data from this list by limiting our analysis to onsets occurring between 19:00 and 05:00 magnetic local time and, 55° and 75° magnetic latitude (the nightside auroral zone). Moreover, we consider substorms where the SML index drops below $-3{,}000$ nT or the SME exceeds 3,000 nT as outliers (334 events) and discard them. As a result, the number of onsets used from the SML list is reduced to ~29,000.

The substorm onset identification algorithm of (Newell & Gjerloev, 2011a) is as follows: The SML data were considered at a 1-min cadence in a sliding 30-min buffer. An onset was identified at t_0 when four conditions are satisfied:

$$SML(t_0 + 1) - SML(t_0) < -15 \quad \text{nT} \tag{1}$$

$$SML(t_0 + 2) - SML(t_0) < -30 \quad \text{nT} \tag{2}$$

$$SML(t_0 + 3) - SML(t_0) < -45 \quad \text{nT} \tag{3}$$

$$\sum_{i=4}^{i=30} SML(t_0 + i)/26 - SML(t_0) < -100 \quad \text{nT} \tag{4}$$

Thus the drop must be sharp (equations (1)–(3), that is, 45 nT in 3 min), and sustained (equation (4), i.e., must be 100 nT below the initial value for the remainder of the half hour on average). The SML onset is then placed at t_0, the last minute before a 15-nT drop. Once an onset is identified, the algorithm advances ahead 20 min, which is thus the minimum permitted time between two consecutive onsets.

In this study we formulate the substorm onset prediction as a binary classification problem. Simply put, our goal is to forecast the occurrence probability of a substorm over the next 60 min, using a 120-min time history of solar wind and IMF parameters (B_x, B_y, B_z, V_x, and N_p). Specifically, each data point used to train the model consists of a 120×5 input matrix of the solar wind and IMF, and an output which is a label indicating the presence (labeled as 1) or absence (labeled as 0) of an onset over the next 60 min. (A 120-min time history was chosen to sufficiently cover the trend in B_z prior to the prediction time.) Figure 1a shows an example of a data point used to train the deep learning model which we present later. The 2-D input array is shaded in gray, and the output binary label is determined based on the presence/absence of a substorm within the next 60 min of the prediction time marked by the first vertical dotted black line from the left. The SML index is shown to confirm the presence of substorm activity, the vertical dotted red line indicates the actual onset time identified using the procedure described in Newell and Liou (2011). A data point is discarded if there are
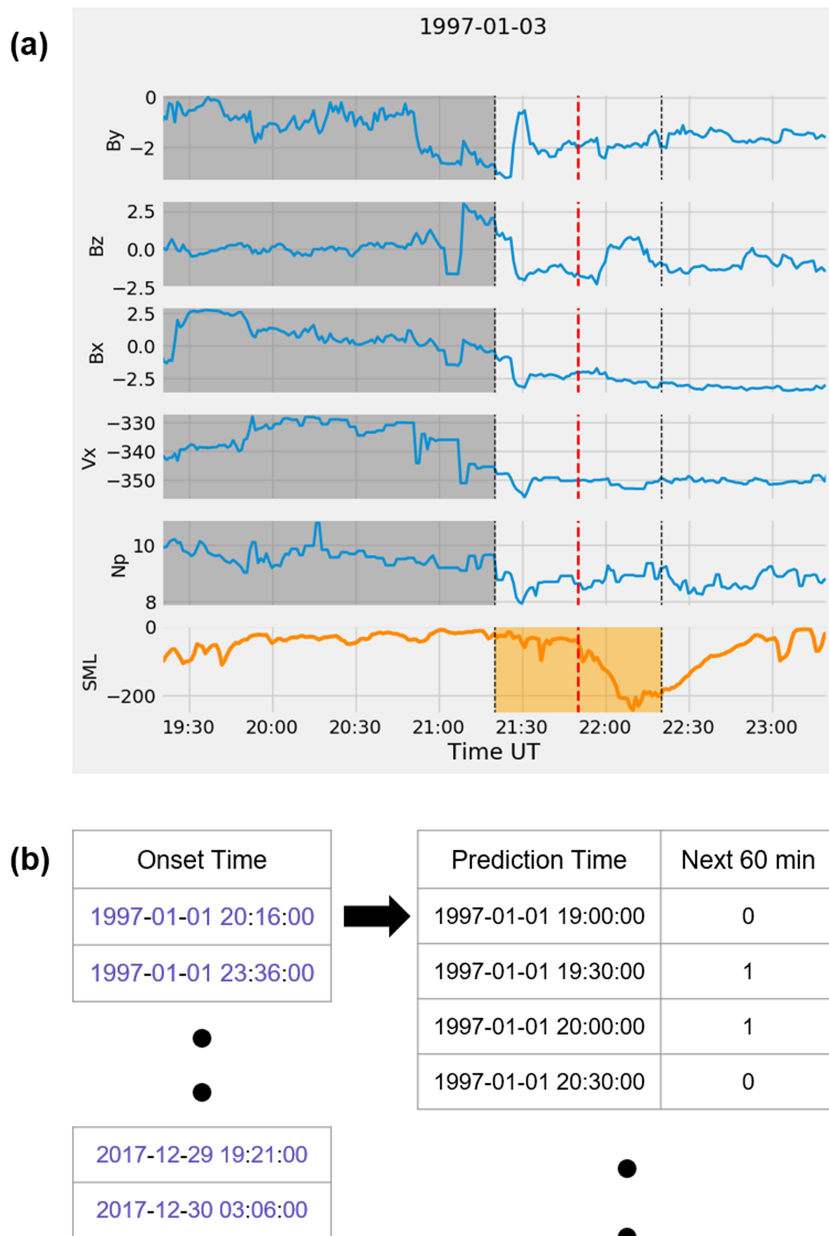
Figure 1. (a) An example of data point classification to a deep learning mode. The 2-D input array is shaded in gray and the output binary label is determined based on the presence/absence of substorm within the next 60 min of the prediction time marked by the first vertical dotted black line from the left. The red dotted line indicates the actual onset time. In this example, the output label is 1, corresponding to presence of substorm onset in the 60-min time interval shaded in orange and marked by the two vertical dotted black lines (b) Conversion of prediction times into binary labels based on the presence/absence of a substorm within the next 60 min.

more than 20 missing values (out of 120 min) in any of the five input parameters, otherwise gaps were filled using linear interpolation. A uniform 10-min delay is added to OMNI solar wind and IMF data to account for the propagation time of solar wind through the magnetosheath, the time for the effect of the external trigger at the magnetopause to affect the magnetotail and trigger substorm onset, and the time for the substorm onset to be detected by ground magnetometers. (Note that we have tried delay times between 0 and 20 min and found our model performance to be robust to uncertainties in delay time estimation). Following a standard practice in machine learning to improve the performance of the model, we normalized each of the input solar wind and IMF variables by subtracting the mean and dividing by its standard deviation,

where the mean and standard deviation were calculated using all the measurements between 1997 and 2017. Figure 1b illustrates the conversion of prediction times, spanning from the year 1997 to 2017 with a step of 30 min, into binary labels based on the presence/absence of a substorm within the next 60 min. This conversion process results in a 1:6 ratio in the number of onsets versus non-onsets. We balanced the class labels by randomly selecting an equal number of non-onset intervals before training the model. The final data set has a total number of 60,678 data points, with 30,339 for each label.

2.2. Technical Definitions

In this section we list technical terms and model performance metrics (e.g., Lee, 2019) that we use in the later parts of this report to evaluate model performance.

True positive (TP) = correctly identified, that is, a substorm onset occurs and the model predicts it correctly.

False positive (FP) = incorrectly identified, that is, a substorm onset does not occur but the model incorrectly predicts it as a substorm onset.

True negative (TN) = correctly rejected, that is, a substorm onset does not occur and the model correctly predicts it as a nonsubstorm event.

False negative (FN) = incorrectly rejected, that is, a substorm onset occurs but the model does not predict it.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

$$\text{Recall (True positive rate)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{False negative rate} = \frac{\text{False Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{F1-Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is the harmonic average of the precision and recall, where an F1-score reaches its best value at 1 (perfect precision and recall) and worst at 0.

2.3. Model Development

In this study, we train a convolution neural network (CNN) model to forecast the occurrence probability of a substorm onset, P_{onset}, over the next 1 hr using a 120-min history of IMF and solar wind parameters. CNNs, first proposed by Cun et al. (1990), are a special kind of multilayer neural network. Unlike conventional machine-learning techniques, which are limited in their ability to process image data in their raw form, CNNs were designed to recognize visual patterns directly from pixel images with minimal preprocessing. They were shown to be very effective in hierarchically capturing complex spatial and temporal relations in the data set (Cun et al., 1990). Since the 2-D input arrays in this problem (see Figure 1a for an example data point) can essentially be treated as 2-D images, CNNs are therefore ideally suited for capturing the dynamics of solar wind and IMF parameters that can eventually drive substorms.

2.3.1. Model Architecture

We developed three CNN models based on the latest variants of CNNs, such as ResNet (residual network; He et al., 2016), FCN (fully convolutional network; Wang et al., 2017), and MLSTM-FCN (multivariate long short-term memory coupled with fully convolutional network; Karim et al., 2018). A recent study (Fawaz et al., 2019) tested nine recently developed deep learning classifiers over 12 multivariate time series classification data sets and ranked ResNet and FCN as the best in terms of average accuracy over 10

runs of each implemented model. After an extensive amount of model tuning we observed that all three variants of the CNNs showed very similar performance; no model had a clear advantage over the others. The ResNet CNN, however, is relatively easier for model tuning and is less prone to adverse effects of adding more layers (Glorot & Bengio, 2010). We therefore decided to adopt the ResNet architecture and analyze its forecasts in greater detail.

Figure 2a shows the schematic of ResNet CNN architecture we have used for our application. The input layer takes a 2-D array with 120 min (time history) × 5 (solar wind and IMF) parameters (see Figure 1a for an example data point). The first two convolution layers (i.e., the layers that have a set of independent filters which perform convolution operations to nodes in preceding layers) and the subsequent max pooling layer (i.e., a downsampled version of the preceding layer) were added to reduce the input dimension while capturing some low-level features from the input data, unlike the implementation of the ResNet model by Fawaz et al. (2019). After the max pooling layer are two Resnet Units, each consists of three identical convolution layers stacked in series. There is an additional "shortcut connection" between the input and output of each ResNet Unit that simply performs identity mapping and its output is added to the output of the stacked layers. Resnet Units serve two purposes: (1) to make the architecture deeper so that the later convolution layers can capture the complex relationship between different input features, and (2) to overcome the problem of vanishing/exploding gradients through the "shortcut connection" (He et al., 2016). The issue of vanishing/exploding gradients arises when training deeper networks because, as the gradients (with respect to the loss function) are back-propagated to earlier layers, repeated multiplication may make the gradients infinitely small/large and thereby prevent the network from training further (Glorot & Bengio, 2010). We tested this ResNet CNN with different numbers of ResNet units, varying between 1 and 10 and found that a CNN model with two ResNet units performs the best on a validation data set (see the next subsection for definition) in terms of F1-score. A max-pooling layer is applied to the output of the second ResNet Unit, which is then flattened to form a long 1-D layer, connected to the output layer after one more 1-D layer. The output layer has two nodes—one for the substorm class and the other for the nonsubstorm class. A SoftMax activation function, a normalized exponential function that maps the preceding vector into a normalized probability distribution, is applied to these two nodes to obtain probability estimates for the two classes. In summary, for any given time instant, the model is designed to accept a 120-min history of the five solar wind and IMF parameters as inputs and output the probability of an onset over the next 60 min. The basic rationale in this ResNet CNN architecture, or CNN in general, is that it first looks for low level features and then gradually builds up more abstract concepts through a series of convolution layers.

2.3.2. Model Training

The ResNet architecture shown in Figure 2a is implemented in Keras (Chollet, 2015) with Tensorflow backend and trained using multi-GPUs on the High-Performance Computing (HPC) system available at Virginia Tech. Each model was trained using a back-propagation algorithm (Rumelhart et al., 1986), a widely used learning algorithm which minimizes the training loss by adjusting the weights according to a gradient descent. The objective of training is to minimize the difference between the output and target pairs (i.e., the loss value) by optimizing the weights of the network. Equation (5) shows the expression of binary cross-entropy loss which our model tries to minimize. The variables y and p in equation (5), respectively, represent the actual label (i.e., 1 for substorm onset or 0 for non-onset) and the substorm onset probability predicted by the model. We used "Adam" optimizer (Kingma & Ba, 2014), which is a method for efficient stochastic gradient descent optimization. The adjustable parameters of "Adam" used in this study are the following: learning rate α=0.00001, learning rate decay = 0.0, exponential decay rates β_1=0.9, and β_2=0.999. See Kingma and Ba (2014) for the definitions of these adjustable parameters. We split the data into three different data sets while keeping the date points in chronological order. Such a split prevents the model from overfitting as the test and validation time periods are completely independent of the training time period: training (70%), validation (15%), and test (15%) data sets. Training data set was used for model training, validation data set was used for model tuning and model selection, and test data set was used for evaluating model performance (Ripley, 1996; Russell & Norvig, 2010). Figures 2b and 2c, respectively, show the loss and accuracy versus number of epochs for the ResNet CNN model with red color for training data and blue for validation data. An Epoch is a complete iteration over the entire training data. Both the training and validation losses sharply decrease in the first 15 epochs and then gradually reach their own asymptotic values; 200 epochs are enough for the loss to converge. More importantly, the loss curves show no sign of overfitting, indicated by
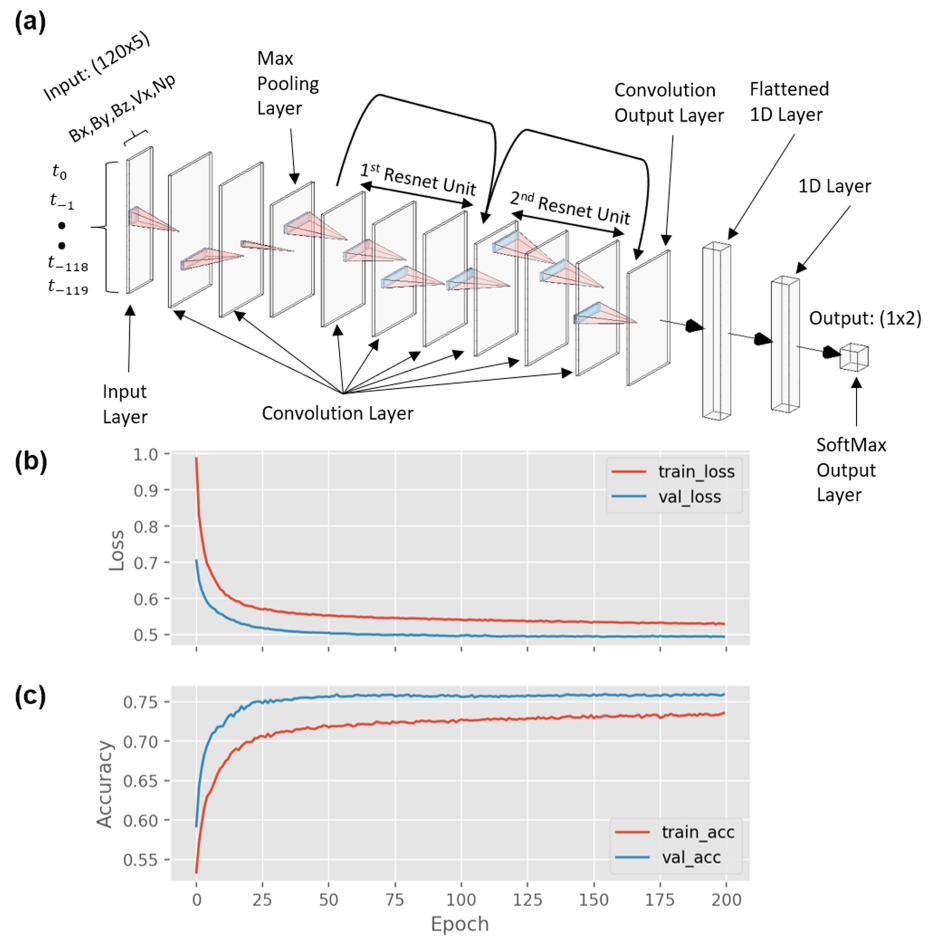
Figure 2. (a) A schematic of ResNet CNN architecture with two ResNet Units for multivariate time series processing of solar wind and IMF data. The input layer takes a 2-D input array with 120×5 elements and the output layer renders probability estimates. The output layer has two nodes, one for the substorm class and the other for the nonsubstorm class. (Note that the layer dimensions are shown in log scale.) Each polyhedron between the adjacent convolution layers represent the mapping of a previous layer to the next through a two-dimensional convolution. (b) Loss and (c) accuracy curves of the model training on training (red curve) and validation (blue curve) data sets. An Epoch is a complete iteration over the entire training data. CNN = convolution neural network.

the loss and accuracy curves for the validation data set not diverging from those for training data set after a certain epoch.

$$Loss = - \quad (y \quad \log(p) + (1-y) \quad \log(1-p)) \qquad (5)$$

3. Results

In the previous sections, we focused on data processing, model architecture, and training. We formulated the problem as a binary classification task and developed a ResNet CNN model that can be used to forecast the onset of a magnetic substorm over the next 60 min, using solar wind and IMF as inputs. In the current section, we present the model forecasts, analyze its performance on a statistical scale, and discuss the role of different input parameters in driving the predictions. It is important to note that all the analysis presented in this section uses the test subset of the data and excludes all intervals from the training and validation subsets.

3.1. Model Predictions

As a first step, we present four example forecasts made by the model in Figure 3; one forecast from each of the TP, FP, FN, and TN categories is shown. The figure presents the inputs (B_y, B_z, B_x, V_x, and N_p) fed to the
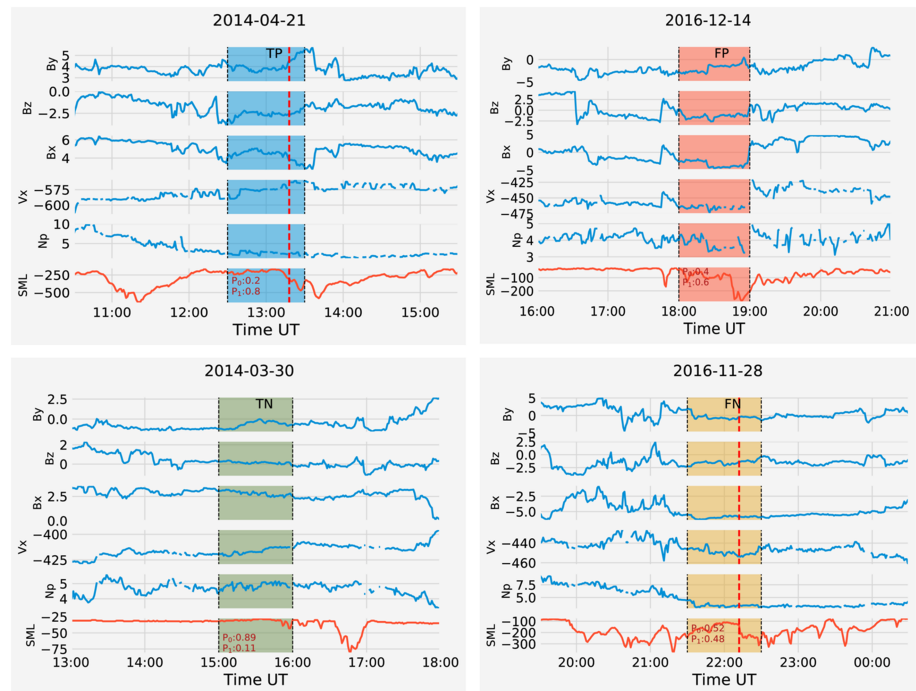
Figure 3. Example model predictions, clockwise from top left are True Positive (TP), (top right) False Positive (FP), (bottom right) False Negative (FN), and (bottom left) True Negative (TN) predictions. In each figure, the top three panels show IMF B_y, B_z, and B_x, respectively, the fourth and fifth panels show the solar wind velocity and number density, and the sixth panel presents the SML index. The dashed black lines and/or the shaded regions mark the prediction interval and the dashed red lines indicate actual onset time provided by SuperMAG. The probabilities associated with each forecast are shown in the bottom panel. See text for details.

model along with SML index which is indicative of substorm activity. The time period shown in each plot is centered around the prediction interval (1 hr) and spans 5 hr, including the 2-hr input period. The prediction interval in each of the plots is indicated by the shaded regions, and the inputs to the model are taken from the 2-hr period preceeding this interval. The actual onset times identified by SuperMAG are indicated by the red dashed lines for the TP and FN categories and the probabilities associated with each forecast are annotated in the bottom panels.

In the TP example, the model forecasted a potential onset for the time period covering 1230 UT and 1330 UT with a high degree of confidence (P_{onset}=0.80) and an onset was indeed identified in SuperMAG SML index ∼0115 UT, confirming the prediction. A drop in SML index further confirms substorm activity during this period. The input to the model are solar wind and IMF parameters between 1030 UT and 1229 UT; B_z was predominantly negative and V_x was ∼−600 km/s during this period. Moving to the FP case, the model forecasts a substorm onset between 1800 and 1900 UT and the interval preceding the prediction (1600 - 1759 UT) was dominated with a weak negative B_z. It is interesting to note that while the onset detection criteria (equations (1)–(4)) do not find any onset in this time period, a closer examination reveals a drop in the SML index, perhaps associated with a minor substorm or a psuedo-onset. This event failed to satisfy all four criteria formulated by Newell and Gjerloev (2011a) and was therefore classified as a non-onset, whereas our model forecast an onset during this interval based on the input solar wind and IMF conditions. This event shows the challenges associated with accurately identifying actual substorm onsets from the SML data set and separating them from pseudo-onsets and minor geomagnetic disturbances. While the SML-based onset list is the most comprehensive and continuous data set available, it has certain limitations and uncertainties which propagate into the model. We must be clear that our model is trained to predict substorm onsets as specified by the criteria of Newell and Gjerloev (2011a). In the FN case, B_z was mostly weak negative with short periods of positive B_z and V_x was ∼−440 km/s during the input interval (1930–2129 UT). While our model failed to forecast an onset detected by SuperMAG ∼2210 UT in this case, it is important to note that

Table 1
Model Performance for Training, Validation, and Test Data Sets

Class label	Precision	Recall	F1-score	# intervals
Prediction for training data				
0 (Nonsubstorm)	0.77	0.70	0.73	21,236
1 (Substorm)	0.72	0.79	0.75	21,236
Prediction for validation data				
0 (Nonsubstorm)	0.74	0.80	0.77	4,496
1 (Substorm)	0.78	0.73	0.75	4,496
Prediction for test data				
0 (Nonsubstorm)	0.74	0.75	0.74	4,607
1 (Substorm)	0.75	0.73	0.74	4,607

P_{onset} was 0.48 indicating higher uncertainty in the model for this prediction. Finally, in the TN example, the model successfully forecasted a period without any substorm activity between 1500 and 1600 UT with a high degree of confidence (P_{onset}=0.11) and the same is confirmed by the SML index. In summary, two important features can be observed from the figure, first, a negative IMF B_z has a significant impact on the forecasts and, second, the prediction probability (indicated by P_{onset}) is higher for successful predictions compared to the FN case. These features will be analyzed further in the next sections.

3.2. Model Performance

We used the model to forecast a large number of intervals between 1997 and 2017 and the evaluation metrics are summarized in Table 1. The "# intervals" column in the table indicates the number of forecast intervals used during training, validation, and testing. For example, 21,236 nonsubstorm and substorm intervals (a total of 42,472 data points) were used to train the model, whereas 4,607 intervals of each class were used for testing. We can note from the table that the model achieves a F1-score of 0.74 for both the classes on the test data. The model evaluation metrics listed in Table 1 are based on setting the P_{onset} threshold to 0.5, that is, we expect the model forecasted a substorm onset if P_{onset}>0.5 and a nonsubstorm otherwise. We will now analyze the impact of systematically varying this threshold between 0 and 1 on the model performance, using a Receiver Operating Characteristic (ROC) curve. The ROC curve is created by plotting the recall against the false positive rate at various threshold levels and serves as a useful tool for evaluating binary classifiers. We plot the ROC curve for our model in Figure 4a, the threshold value of 0.5 is marked by the "x" symbol. As we increase the threshold, the recall increases; however, this is accompanied by an increase in the false positive rate. There are certain advantages and disadvantages of choosing a certain threshold, and the choice is therefore best left to the end user. In comparison, the ROC curve of a model that makes a prediction based on a random guess is indicated by the diagonal dotted black line. In other words, the further the solid black curve is from the diagonal line, the better the model is at discriminating between onsets and non-onsets. Overall, our model achieves a ROC area of 0.82, which is a measure of the separability between the onset and non-onset classes.

After training, our model achieves an accuracy of ~75% and the remaining 25% of the events are misclassified. Now we separately examine the distributions of P_{onset} for both the onset and non-onset classes to further understand model performance. Such an analysis will provide insights into the confidence of a forecast. For example, a higher P_{onset} (> 0.85) shows that the model is very confident of the forecast compared to a lower value. Figure 4b shows histograms of P_{onset} for both the onset (purple) and non-onset (green) classes. In this figure, we use a threshold of 0.5 (dashed black line) to forecast an onset. So for the non-onset class (i.e., no subsequent substorm) when P_{onset}<0.5, the model's forecast is correct and is categorized as a TN event. Likewise, when $P_{onset} \geq 0.5$ for an onset class (i.e., subsequent substorm) it is categorized as a TP event. On the other hand, P_{onset}<0.5 for an onset class makes the event FN whereas P_{onset}>0.5 for a non-onset class make it FP. All four categories are marked in the figure. We note that the model can predict large number of substorm and nonsubstorm events with high confidence as reflected by the two peaks in the purple (P_{onset}>0.8) and green curves (P_{onset}< 0.15), respectively. However, we find that there are a considerable number of nonsubstorm events misclassified as onsets (FPs) by the model (green curve in regions where P_{onset}>0.5) and a similar albeit less prominent feature is observed for FNs. We will further analyze this behavior in later sections and analyze the origin of this problem in greater detail.

We have so far examined the overall performance of the model forecasts and now we analyze its dependence on the change in geomagnetic activity, binned by the maximum drop in SML index during the prediction horizon interval of 60 min (Δ_{SML}). In Figure 5, we show the number of forecasts (panel a) and P_{onset} (panel b), individually for each prediction category (TP, FP, TN, and FN) for different Δ_{SML} bins. It can be seen that when the maximum drop in SML during the 1-hr forecast interval is less than 100 (corresponding to the −100 to 0 bin), the maximum forecasts belong to the TN category, whereas a noticeable number from the FP and TP categories are observed as well. Furthermore, the number of TPs increases while the number of TNs decreases with Δ_{SML}. This higher number of TNs in the first bin can be attributed to the fact that lower Δ_{SML} periods are usually quiet with very few onsets, and the increase in the number of onsets with Δ_{SML}
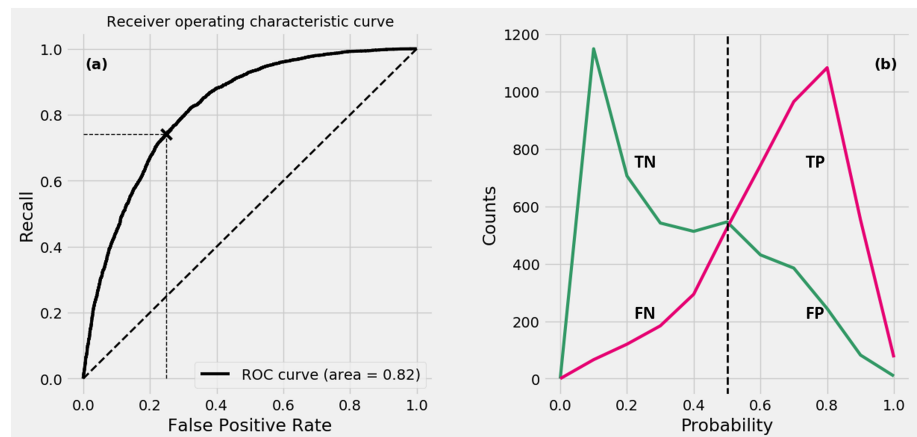
Figure 4. (a) Receiver operating characteristic (ROC) curve for the forecasts. The 0.5 threshold is marked by the "x" sign. The dotted blue line represents the ROC curve of a model that makes a prediction based on a random guess. (b) Histograms of predicted probability of substorms, P_{onset}, forecasted for both onset (maroon) and non-onset (green) classes.

accounts for the higher TPs in those bins. Panel (b) shows a plot of P_{onset} which is indicative of the confidence assigned to the forecast. A very high (low) P_{onset} suggests the model is confident in its forecast of an onset (non-onset). However, an intermediate value of P_{onset} between 0.35 and 0.65 typically is indicative of increased uncertainty in the predictions. It can be concluded from the figure that uncertainty in the model while forecasting onsets (non-onsets) is higher when $|\Delta_{SML}|$ is low (high). Another key feature to note is that the maximum number of FPs occur when the drop in SML is less than 100 nT, and moreover, the model is not very confident in predicting onsets in this range. We will analyze this behavior further and discuss its impact on model predictions in the discussion section.

3.3. Overlap in the Input Parameters

We now examine the behavior of input parameters to understand their influence on model performance. Of particular interest is IMF B_z which is expected to have the maximum impact on substorm onsets (Caan et al., 1977; Lyons, 1995; Lyons et al., 1997). In Figure 6, we present a plot of B_z with respect to the prediction time for each of the four prediction categories (TP, FP, FN, and TN), in other words, ΔT_{pred} indicates the time IMF
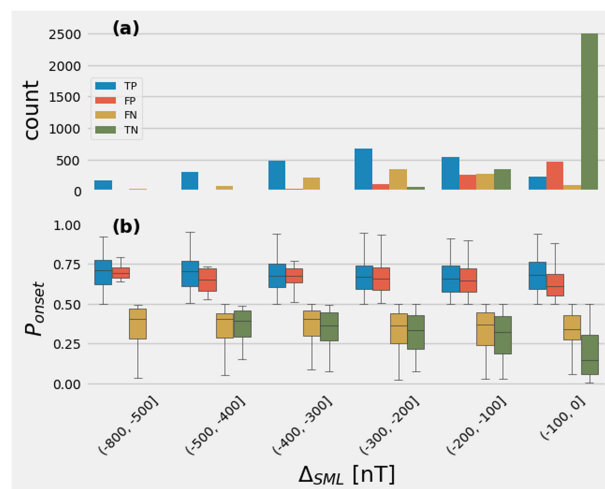

Figure 5. Variability in model predictions with Δ_{SML}. (a) Distribution of all the four forecast categories (TP, FP, FN, and TN) versus Δ_{SML} bins, and (b) variability of P_{onset} for the four forecast categories versus Δ_{SML}. Each box presents the quartiles (minimum, first quartile, median, third quartile, and maximum) of P_{onset} and values beyond ± 1.5 times (interquartile range) are discarded as outliers.
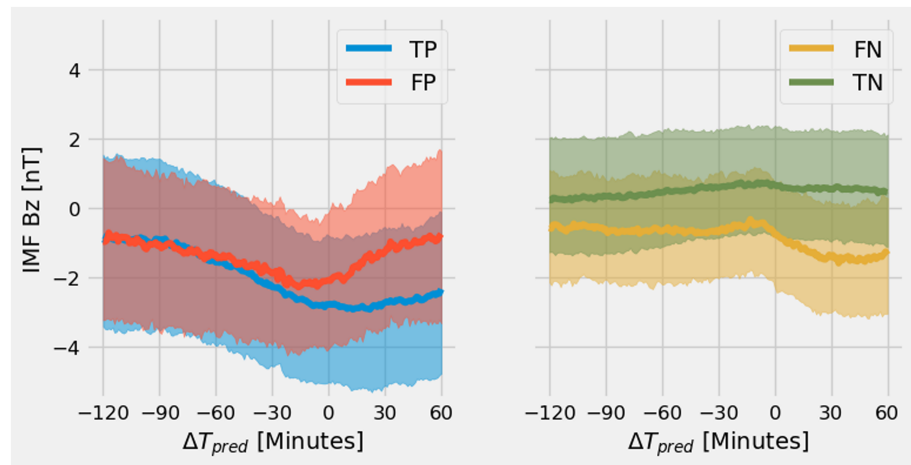
Figure 6. Variability in IMF B_z with respect to prediction time for the four forecast categories. Solid lines indicate median values and the shaded region indicates the 25th to 75th percentile values. Left panel shows true positive (blue) and false positive (orange) categories, and right panel shows the true negative (green) and false negative categories (yellow). See text for details.

B_z was measured relative to the prediction time. For example, if the prediction time is 1200 UT, then ΔT_{pred} is −60 min for a B_z value at 1100 UT. The solid lines in the figure indicate median values whereas the shaded regions mark the 25th and 75th percentile range of IMF B_z. While there is significant variance in B_z for all the four forecast categories, certain strong patterns emerge for each of them. It can be noted that the input IMF B_z turns increasingly southward for the TP category as one moves closer to the prediction time (ΔT_{pred}=0). In contrast, IMF B_z turns increasingly northward closer to the prediction time for the TN category. Clearly, there is a sharp distinction in the behavior of B_z between the TP and TN categories and such distinguishing features are expected to improve the model performance. It is further interesting to note that the trend in IMF B_z for the FP (FN) categories is similar to the TP (TN) categories, albeit slightly weaker. This suggests that our model associates onsets with a preceding interval of southward IMF. More importantly, the similarity in IMF B_z trend between TP and FP (as well as TN and FN) categories suggests that the errors in model prediction could be associated with an overlap in input parameters between the onset and non-onset classes. We will now explore the degree of overlap in the input IMF and solar wind parameters further through Principal Component Analysis (PCA).

Each class label has an associated 2-D array with 600 elements (120 by 5) and it is not possible to fully visualize such high dimensional data. We therefore use PCA techniques to reduce the dimensions of the 2-D input arrays, while preserving as much variation in the data as possible, and visualize them in a new coordinate system as shown in Figure 7. Panel (a) shows percent explained variance versus number of principal components and Panels (b)–(d) show the two classes projected onto the first three principal components. In the transformed coordinate system, the first three principal components alone can explain about 65% of the variance in the original input data. As seen from panels (b)–(d), the two classes are strongly overlapped with each other in the reduced 3-D space. Admittedly, we cannot rule out the possibility that the two classes could be more separable through the other principal components that contain the remaining 35% variance. However, the significant overlap between the two classes suggests that large number of substorm and nonsubstorm events have very similar solar wind and IMF conditions in their 120 min of time history intervals prior to the prediction times. This is also apparent from the strong overlap in IMF B_z for the four forecast categories shown in Figure 6.

The strong overlap revealed by the PCA shown in Figure 7 is for all substorm categories. However, we found that the amount of overlap varies when the substorm level is considered—specifically, there is less overlap between substorm and nonsubstorm events when smaller substorms, as characterized by Δ_{SML}, are discarded (results not shown). Here we show how the deep learning model performs for different levels of substorms. Intuitively we would expect the larger substorms to be detected more easily than the smaller ones. Figure 7e shows the recall versus substorm level thresholded by Δ_{SML}. As expected, the model
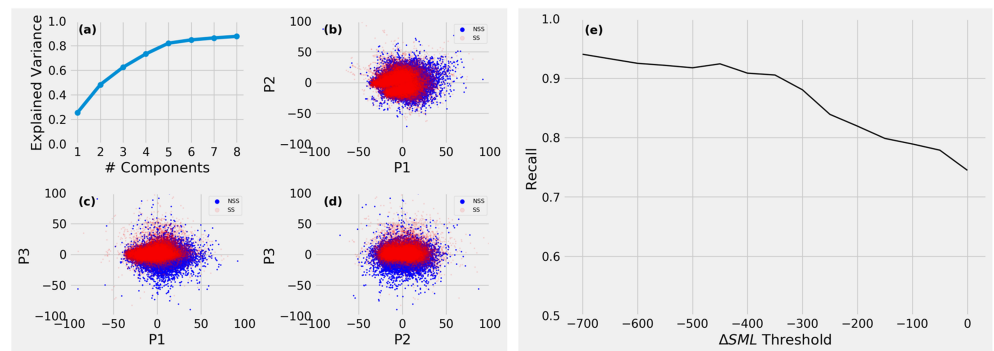
Figure 7. Principal component analysis of the input 2-D arrays of substorm and nonsubstorm events: (a) Percent explained variance versus number of principal components, (b) scatter plot of first two principal components, (c) scatter plot of first and third principal components, (d) scatter plot of second and third principal components, and (e) variation in recall versus substorm levels thresholded by Δ_{SML}.

performance improves for increasing substorm levels. For example, the recall is about 75% when all substorms are considered and it reaches ~95% for substorms with $\Delta_{SML} \geq 700$. The decrease in recall for smaller substorms correlates with more PCA overlap between the two classes (there are a few potential explanations for this finding. See section 4.4 for a detailed discussion on the impact of solar wind/IMF overlap and other factors on model performance).

Table 2
Parameter Ranking

	Precision	Recall	F1 − Score
Bx, By, Bz, Vx, Np	0.75	0.73	0.74
Bx, By, Bz, Vx	0.75	0.71	0.73
By, Bz, Vx, Np	0.74	0.73	0.74
Bx, Bz, Vx, Np	0.74	0.73	0.74
Bx, By, Bz, Np	0.73	0.67	0.7
Bx, By, Vx, Np	0.69	0.58	0.63
By, Bz, Vx	0.75	0.7	0.72
Bz, Vx	0.75	0.7	0.72
Bz	0.71	0.68	0.7
Vx	0.69	0.51	0.58
By	0.64	0.49	0.55
Bx	0.65	0.43	0.52
Np	0.58	0.35	0.44

Note. Precision and recall of onsets and non-onsets when different IMF and solar wind parameters are given as inputs to the model.

3.4. Input Parameter Importance

Deep learning models generally perform better than traditional machine learning models when the data set is large and the relationship between the input data and output labels are complex. However, deep learning models, or any artificial neural networks in general, are known to be difficult for human interpretation because the features generated in the last hidden layer (i.e., the layer just before the output) cannot be directly related to the input features. In our application it is important to have some type of model interpretability which we can relate to our physical understanding of drivers of substorm onsets. To overcome the black box nature of deep learning models, we take the backward stepwise feature selection procedure which is often used in linear regression to rank the input parameter importance. The assumption here is that if a certain input parameter is important in triggering substorms, it should also be an important feature for deep learning models to predict substorm onset. Such an approach was previously used by Weigel et al. (2003) to determine the importance of different parameters in predicting ground magnetic field and its time derivative. Table 2 lists the performance metrics (precision, recall, and F1-score) of the ResNet CNN model for different combinations of input parameters. We can see that B_z and V_x are the most important factors on the model performance such that B_z influences both precision and recall while V_x influences mostly the recall. This is apparent when each of the five features is used as input individually or is dropped from the full model that uses all the five features. The implications of the parameter importance are discussed in more detail later.

3.5. Baseline Comparison

We compare the performance of our model with the methodology described in Lyons et al. (1997) that nowcast substorm onset. A rigorous set of quantitative rules (listed below) were proposed to determine if a substorm is triggered at a given time (t_{on}) based on the variations in IMF B_z 30 min prior to the onset (i.e.,$t_{on}-30$ to t_{on}).

$$N_{B_Z(t_{on}-30:t_{on})<0}\geq22 \tag{6}$$

$$B_Z(t_{on} + \Delta_t)-B_Z(t_{on})\geq0.375 \quad nT/min \tag{7}$$

$$\frac{B_Z(t_{on} : t_{on} + 10\Delta_t)}{10\Delta_t}\geq0.175 \quad nT/min \tag{8}$$

$$B_Z(t_{on} + \Delta_t : t_{on} + 3\Delta_t)\geq B_Z(t_{on}) + 0.15 \quad nT \tag{9}$$

$$B_Z(t_{on} + 3\Delta_t : t_{on} + 10\Delta_t)\geq B_Z(t_{on}) + 0.45 \quad nT \tag{10}$$

Equations (6) through (10) represent the criteria proposed by Lyons et al. (1997). Equation (6) quantifies the criterion that IMF must be southward for at least 22 of the preceding 30 min, representing the growth phase requirement. The remaining equations (7) to (10) quantify the criteria for triggering the onset, a rapid and sustained northward turning of IMF B_z. We find that the Lyons et al. (1997) model can correctly identify ~21% of the substorms listed in the SuperMAG database. Previous studies (Morley & Freeman, 2007; Newell & Liou, 2011; Wild et al., 2009) applied the same criteria on other data sets such as those from POLAR UVI (Liou, 2010) and IMAGE (Frey et al., 2004) and found that all the criteria are satisfied only by a small subset (<30%) of substorm events, similar to our study. It should be emphasized that the Lyons et al. (1997) model was developed for a different purpose, namely, to study the triggering mechanisms of substorms and it has since then been extremely valuable in improving our understanding of substorms. However, from a space weather prediction perspective, there are very few tools/methods available to forecast the onset of a substorm (Tanaka et al., 2015) and we believe our model will serve this purpose.

3.6. Solar Wind-Magnetosphere Coupling Function

While it is difficult to interpret the features captured by any deep learning-based model, modulating the input features and understanding their influence on the prediction can provide some useful insights. Next, we analyze the impact of using different solar wind–magnetosphere coupling functions as inputs to our model.

Table 3 shows performance metrics of the model (precision, recall, and F1-score) when different coupling functions with 120 min of time history are given as model inputs. We found that using $d\Phi/dt$ which is indicative of the rate magnetic flux opened at the magnetopause (Newell et al., 2007) as input to our model gives

Table 3
Precision, Recall, and F-Score of the Model When Different Coupling Functions Are Given as Model Inputs

Model inputs	Precision	Recall	F1-score
$V_x B_T$	0.67 ± 0.03	0.65 ± 0.06	0.65 ± 0.03
$E_{KL} = V_x B_T \sin^2(\frac{\theta_c}{2})$ (Kan & Lee, 1979)	0.71 ± 0.03	0.71 ± 0.05	0.71 ± 0.01
$E_{TL} = N_p^{\frac{1}{2}} V_x^2 B_T \sin^6(\frac{\theta_c}{2})$ (Temerin & Li, 2006)	0.72 ± 0.03	0.68 ± 0.04	0.70 ± 0.01
$\frac{d\Phi_{MP}}{dt} = V_x^{\frac{4}{3}} B_T^{\frac{2}{3}} \sin^{\frac{8}{3}}(\frac{\theta_c}{2})$ (Newell et al., 2007)	0.72 ± 0.03	0.75 ± 0.04	0.73 ± 0.01
B_x, B_y, B_z, V_x, N_p	0.72 ± 0.02	0.77 ± 0.04	0.74 ± 0.01
$B_x, B_y, B_z, V_x, N_p, \frac{d\Phi_{MP}}{dt}$	0.72 ± 0.02	0.78 ± 0.04	0.74 ± 0.01

Note. The value of each metric is given in the form of mean–std., where the mean and standard deviation are computed by training the model repeatedly on 10 different training/test/validation sets.

the best performance in predicting substorm onsets (F1-score = 0.73) compared to other coupling functions. The most successful implementation of the model, which uses the raw solar wind and IMF features (i.e., B_x, B_y, B_z, V_x, N_p) as input features, achieves an F1-score of 0.74, and thus performs slightly better than the other models that use coupling functions as features. Furthermore, providing $d\Phi/dt$ as input to the model in addition to the raw solar wind and IMF features doesn't improve the model performance any further. These results suggest that our model is able to replicate the behavior of the coupling functions which quantify certain aspects of solar wind - magnetosphere interactions, from the input solar wind and IMF parameters.

4. Discussion

In this study, we presented the first deep learning-based approach to forecast the occurrence probability of substorm onsets over the next 60 min. The model was shown to achieve 72±2% precision and 77±4% recall rates and an ROC area of 0.83 (Figure 4a). We further showed that a majority of the false positives are associated with small substorms ($|\Delta_{SML}|$<100 nT). The errors in model prediction can be attributed to a strong overlap in input parameters between the onset and non-onset classes as demonstrated by PCA in Figure 7. Finally, we found that B_z and V_x have the most significant impact on model performance. In this section, we further analyze our findings and compare them with observations from previous studies.

4.1. Comparison With Previous Approaches

Some studies used neural networks to predict auroral indices over the next few minutes using solar wind and IMF as inputs (Gavrishchaka & Ganguli, 2001; Gleisner & Lundstedt, 1997; Hernandez et al., 1993), and provided a mechanism to forecast substorm. While these studies serve as excellent starting points, there are uncertainties and inaccuracies associated with using auroral indices (such as AE) as a proxy for identifying substorm activity (Kullen et al., 2009). More recently, Luo et al. (2013) developed empirical models of auroral indices and showed that predicting short term variations is more challenging than predicting the long term average trends. Furthermore, Barkhatov et al. (2017) used an Elman artificial neural network to make a short-term forecast of the AL index, and showed that taking into account the 2- to 3-hr history of solar wind kinetic energy loading the magnetosphere significantly improves the accuracy of the model. In this study, we build on these approaches, by developing a model that directly forecasts the occurrence probability of a substorm onset, with a higher prediction horizon of 60 min. In agreement with these studies, we find that incorporating the history of solar wind and IMF into the model, indicating the growth phase of the substorm, is crucial for making accurate forecasts. Furthermore, our analysis shows that the deep learning-based approach is capable of capturing complex interactions between the solar wind and magnetosphere and can serve as an alternative to those approaches that are based on coupling functions. These results are supported by an early study (Weigel et al., 2003), which successfully showed that a neural network can be used to develop a "data-derived" nonlinear coupling function to predict ground magnetic field and its time derivative at different local times using solar wind and IMF over a 3-hr period as inputs.

4.2. External Triggering of Substorms

The epoch analysis presented in Figure 6 shows that a crucial factor driving the model's predictions is the trend in IMF B_z. If B_z starts turning increasingly negative ∼60 min prior to the prediction time, the model is more likely to forecast the occurrence of a substorm onset. This behavior suggests the model is capturing the growth phase of the substorm during which energy from the solar wind is accumulated in the magneto-tail, a widely accepted condition necessary for triggering a substorm (Freeman & Morley, 2009; Johnson & Wing, 2014; Henderson et al., 1996; Lyons, 1995; Lyons et al., 1997; McPherron et al., 1973; Russell, 2000).

A few studies (Caan et al., 1977; Lyons, 1995; Lyons et al., 1997; Russell, 2000) postulated another condition necessary to trigger a substorm, namely, a sharp northward turning of IMF B_z close to or coinciding with the onset. However, the ability of our model to accurately forecast (up to 1 hr ahead) substorm onset without any IMF or solar wind input close to the actual onset time strongly suggests that the second condition may not always be necessary to trigger an onset. Moreover, the epoch analysis doesn't show a strong trend indicative of IMF northward turning immediately prior to the onset. Our findings suggest that the behavior of a majority of substorm onsets is in agreement with another set of studies (Freeman & Morley, 2009; Johnson & Wing, 2014; Henderson et al., 1996; Morley & Freeman, 2007; Newell & Liou, 2011), which argued that initial elevated solar wind driving conditions is the only necessary condition for triggering a substorm, and the northward turning of IMF B_z is coincidentally rather than causally related to onsets (Freeman & Morley, 2009; Morley & Freeman, 2007).

4.3. Input Parameter Importance

Determining the roles of different IMF and solar wind parameters in triggering a substorm has been a subject of interest for several decades (Caan et al., 1977; McPherron et al., 1973; Newell et al., 2016; 2013; Petrukovich et al., 2000; Troshichev et al., 1986). In this study, we modulated the inputs given to our machine learning model to quantify the influence of different parameters in forecasting substorm onsets. We found IMF B_z, V_x, and IMF B_y had the most significant influence on the forecasts (in the same order), and including them as inputs improved the accuracy of the forecasts whereas dropping them reduced it, as shown in Table 2.

A southward IMF B_z was expected to be a significant factor because the magnetic flux accumulation that occurs during substorm growth phase via dayside magnetic reconnection is well known to occur during southward IMF B_z. In our analysis IMF B_y ranked third in importance. While not as impactful as B_z, under dominant B_y conditions open flux can still be added to the magnetosphere via dayside reconnection (Sandholt et al., 1998) and thus contribute to the growth phase of a substorm. Several earlier studies have confirmed the role of B_y in driving substorms, especially when B_z is positive and cannot contribute to the growth phase significantly (Lee et al., 2010; Petrukovich et al., 2000; Troshichev et al., 1986; Wild et al., 2009).

Our model ranks V_x as the second most influential factor in forecasting substorms. This is in good agreement with (Li et al., 2007; Luo et al., 2013; Newell et al., 2016) who showed that while the B_z component of IMF plays an important role in modulating magnetospheric dynamics, the role of solar wind velocity is more important than is usually assumed. Furthermore, we confirm the result reported by Li et al. (2007) that number density by itself is the least significant factor influencing substorm electrodynamics (Table 2). Recent observations (Newell et al., 2013; Tanskanen, 2009) also showed that higher V_x is directly correlated with increased substorm activity. Moreover, V_x has been a component of many solar wind-magnetosphere coupling functions that estimate the dayside magnetopause merging, in fact, 17 out of 20 coupling functions listed in (Newell et al., 2007) include V_x. This suggests higher V_x can increase the open flux transportation to the magnetotail, and thus, facilitate substorm growth phase. Beyond the role of V_x in modulating dayside merging, MHD simulation studies of (Goodrich et al., 2007; Pulkkinen et al., 2007) show that higher solar wind speed makes the magnetotail much more unstable, which may also trigger more substorms.

4.4. Impact of Solar Wind/IMF Overlap on Model Performance

An important observation from Figure 5 is that the forecasts belonging to the false positive and true negative classes are most apparent when $|\Delta_{\text{SML}}|$ is less than 200 nT, when the geomagnetic activity is relatively low. However, during periods of dynamic geomagnetic activity (higher $|\Delta_{\text{SML}}|$), the number of TN and FP forecasts drops significantly and the TP forecasts dominate over the other classes. In other words, the FP

forecasts occur during similar conditions as the TN forecasts and this behavior can be attributed to the degree of overlap between the onset and non-onset classes (Figure 7a). While the lower Δ_{SML} bins are predominantly associated with quiet conditions (no onsets) they still had a notable number of "minor" substorms or pseudo-onset, increasing the number of FPs and the uncertainty in the model. One such example is seen in the top-right panel of Figure 3 where our model predicts an onset but the SML index only shows a small disturbance or perhaps a pseudo-onset. This uncertainty in forecasting "minor" substorms is further confirmed by a drop in P_{onset} (Figure 5) when $|\Delta_{SML}| < -200$ nT as well as the increase in recall with $|\Delta_{SML}|$ threshold (Figure 7b).

The performance of our model relies on the assumption that a recent history of solar wind and IMF can determine the dynamics of magnetic substorms. However, the distinction in solar wind and IMF conditions prior to nonsubstorm intervals, pseudo-onset and minor substorms is not sharp (Kullen & Karlsson, 2004). The lack of a clear and sharp differentiating factor in solar wind and IMF prior to the two classes (substorm and nonsubstorm) indicates the existence of an overlap between the classes (observed in the PCA analysis shown in Figure 7) and such an overlap can reduce the performance of any machine model (Batista et al., 2004). There are also other assumptions pertaining to the input data and the model architecture, violation of which will affect the mode performance. For example, similar to other approaches such as Magnetohydrodynamics (MHD) and time series based models, our model is developed on the assumption that an upstream point measurements can fully represent the solar wind and IMF conditions across the entire magnetopause. However, such a situation might not be completely valid because of the uncertainties in point measurements (Petrukovich et al., 2000). Another assumption is that the ResNet model will capture all of the dynamics of the magnetosphere. In principle, a neural network could fully capture the dynamics, but this may not be possible in practice. One motivation for using a deep neural network, such as the one in this study, is that it provides better performance than a shallow neural network even though both are "universal approximators" (Cybenko, 1989; Hornik et al., 1989). It is also important to note that the performance of our model is also coupled with the substorm onset list derived using the SML index and any uncertainties or shortcomings in the list may propagate into the model. Nonetheless, the SML based onset list is the most comprehensive and continuous data set available, and is ideally suitable for training a deep learning model.

While a majority of substorms are associated with southward IMF B_z or increased solar wind driving (Kullen & Karlsson, 2004; Lyons, 1995; McPherron et al., 1973; Newell et al., 2016), several studies have reported observations of substorms even during northward B_z intervals (Lee et al., 2010; Miyashita et al., 2011; Petrukovich et al., 2000). The presence of such intervals, albeit in fewer number, further reduces the model performance. Such substorms were suggested to be driven by factors not related to recent history of solar wind and IMF. For example, magnetosphere can take several hours to release the excess energy which may have been supplied during a previous southward IMF interval (Akasofu, 1975; Kullen & Karlsson, 2004; Lee et al., 2010; Miyashita et al., 2011). In such cases, the 120 min of time history is not enough and considering longer time history may help fully capture the state of the system. Another factor is the small-scale spatial variability in IMF during weak solar wind driving conditions which may lead to small regions of southward IMF that can supply energy through localized reconnection regions (Kullen & Karlsson, 2004; Petrukovich et al., 2000).

In summary, we find that while the recent history of solar wind and IMF can be used to forecast a majority substorm onsets, the significant overlap in solar wind and IMF conditions prior to minor substorms and non-onsets reduces the model performance. Other factors, such as the implicit model assumptions and the lack of information on the internal state of the magnetosphere (McPherron et al., 2008), may be additional factors that reduce the model performance.

5. Conclusions and Summary

In this paper, we presented the first deep learning-based approach to forecast the onset of a magnetic substorm, as identified by the SuperMAG SML index. A ResNet convolutional neural network architecture was developed to predict the occurrence probability of an onset over the next 1 hr by using the 120-min history of solar wind and IMF parameters (V_x, N_p, B_x, B_y, and B_z) prior to onset as inputs to the model. The model was trained on a data set consisting of substorm onset intervals derived from the list of onsets compiled by SuperMAG between 1997 and 2017, as well as an equal number of non-onset intervals during the

same time period. We further split the entire data into three subsets: training (70%), validation (15%) and test (15%) and the model achieved 72±2% precision and 77±4% recall rates on the test set. These precision and recall rates show that our model is able to accurately forecast the occurrence of a majority of substorm onsets without any solar wind or IMF input close to the actual onset time. This observation suggests that the majority of substorms are not externally triggered by B_z northward turning. To analyze the influence of different parameters on the forecasts, we modulated the input IMF and solar wind parameters provided to the model and found that IMF B_z and solar wind velocity (V_x) have the most significant influence and provide maximum predictive power. Finally, principal component analysis reveals that small substorms share very similar solar wind and IMF time history conditions with nonsubstorms, suggesting the inputs provided to the model may not be sufficient by themselves to forecast all substorms, and so other factors such as internal magnetospheric instability are at play.

References

Akasofu, S.-I. (1964). The development of the auroral substorm. *Planetary and Space Science*, *12*(4), 273–282. https://doi.org/10.1016/0032-0633(64)90151-5

Akasofu, S.-I. (1975). The roles of the north-south component of the interplanetary magnetic field on large-scale auroral dynamics observed by the dmsp satellite. *Planetary and Space Science*, *23*(10), 1349–1354. https://doi.org/10.1016/0032-0633(75)90030-6

Akasofu, S.-I., Chapman, S., & Meng, C.-I. (1965). The polar electrojet. *Journal of Atmospheric and Terrestrial Physics*, *27*(11), 1275–1305. https://doi.org/10.1016/0021-9169(65)90087-5

Angelopoulos, V. (2008). The themis mission. *Space Science Reviews*, *141*(1), 5. https://doi.org/10.1007/s11214-008-9336-1

Angelopoulos, V., McFadden, J. P., Larson, D., Carlson, C. W., Mende, S. B., Frey, H., et al. (2008). Tail reconnection triggering substorm onset. *Science*, *321*(5891), 931–935. https://doi.org/10.1126/science.1160495

Baker, D. N., Pulkkinen, T. I., Angelopoulos, V., Baumjohann, W., & McPherron, R. L. (1996). Neutral line model of substorms: Past results and present view. *Journal of Geophysical Research*, *101*, 12,975–13,010. https://doi.org/10.1029/95JA03753

Barkhatov, N. A., Vorob'ev, V. G., Revunov, S. E., & Yagodkina, O. I. (2017). Effect of solar dynamics parameters on the formation of substorm activity. *Geomagnetism and Aeronomy*, *57*(3), 251–256. https://doi.org/10.1134/S0016793217030021

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter ACM stands for Association for Computing Machinery SIGKDD stands for Special Interest Group on Knowledge Discovery and Data Mining SIGKDD Explor Newsl*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Belakhovsky, V., Pilipenko, V., Engebretson, M., Sakharov, Y., & Selivanov, V. (2019). Impulsive disturbances of the geomagnetic field as a cause of induced currents of electric power lines. *Journal of Space Weather and Space Climate*, *9*(A18), 19. https://doi.org/10.1051/swsc/2019015

Caan, M. N., McPherron, R. L., & Russell, C. T. (1977). Characteristics of the association between the interplanetary magnetic field and substorms. *Journal of Geophysical Research*, *82*(29), 4837–4842. https://doi.org/10.1029/JA082i029p04837

Chollet, F., and others (2015). Keras. https://keras.io

Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., & Henderson, D. (1990). Advances in neural information processing systems 2, *chap Handwritten Digit Recognition with a Back-propagation Network* (pp. 396–404). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Davis, T. N., & Sugiura, M. (1966). Auroral electrojet activity index ae and its universal time variations. *Journal of Geophysical Research (1896-1977)*, *71*, 785–801. https://doi.org/10.1029/JZ071i003p00785

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*, 917–963.

Freeman, M. P., Forsyth, C., & Rae, I. J. (2019). The influence of substorms on extreme rates of change of the surface horizontal magnetic field in the u.k. *Space Weather*, *17*, 827–844. https://doi.org/10.1029/2018SW002148

Freeman, M. P., & Morley, S. K. (2004). A minimal substorm model that explains the observed statistical distribution of times between substorms. *Geophysical Research Letters*, *31*, L12807. https://doi.org/10.1029/2004GL019989

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*, L21101. https://doi.org/10.1029/2009GL040621

Frey, H. U., Mende, S. B., Angelopoulos, V., & Donovan, E. F. (2004). Substorm onset observations by image-fuv. *Journal of Geophysical Research*, *109*, A10304. https://doi.org/10.1029/2004JA010607

Gavrishchaka, V. V., & Ganguli, S. B. (2001). Optimization of the neural-network geomagnetic model for forecasting large-amplitude substorm events. *Journal of Geophysical Research*, *106*, 6247–6257. https://doi.org/10.1029/2000JA900137

Gjerloev, J. W. (2012). The supermag data processing technique. *Journal of Geophysical Research*, *117*, A09213. https://doi.org/10.1029/2012JA017683

Gleisner, H., & Lundstedt, H. (1997). Response of the auroral electrojets to the solar wind modeled with neural networks. *Journal of Geophysical Research*, *102*, 14,269–14,278. https://doi.org/10.1029/96JA03068

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (vol. 9, pp. 249–256). Chia Laguna Resort, Sardinia, Italy: PMLR. http://proceedings.mlr.press/v9/glorot10a.html

Goodrich, C. C., Pulkkinen, T. I., Lyon, J. G., & Merkin, V. G. (2007). Magnetospheric convection during intermediate driving: Sawtooth events and steady convection intervals as seen in lyon-fedder-mobarry global mhd simulations. *Journal of Geophysical Research*, *112*, A08201. https://doi.org/10.1029/2006JA012155

Gubby, R., & Evans, J. (2002). Space environment effects and satellite design. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(16), 1723–1733. space Weather Effects on Technological Systems https://doi.org/10.1016/S1364-6826(02)00122-0

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, *2016*, 770–778. https://doi.org/10.1109/cvpr.2016.90

Henderson, M. G., Reeves, G. D., Belian, R. D., & Murphree, J. S. (1996). Observations of magnetospheric substorms occurring with no apparent solar wind/imf trigger. *Journal of Geophysical Research*, *101*, 10,773–10,791. https://doi.org/10.1029/96JA00186

Hernandez, J. V., Tajima, T., & Horton, W. (1993). Neural net forecasting for geomagnetic activity. *Geophysical Research Letters*, *20*, 2707–2710. https://doi.org/10.1029/93GL02848

Horne, R. B., Glauert, S. A., Meredith, N. P., Boscher, D., Maget, V., Heynderickx, D., & Pitchford, D. (2013). Space weather impacts on satellites and forecasting the earth's electron radiation belts with spacecast. *Space Weather*, *11*, 169–186. https://doi.org/10.1002/swe.20023

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

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

Kan, J. R., & Lee, L. C. (1979). Energy coupling function and solar wind-magnetosphere dynamo. *Geophysical Research Letters*, *6*, 577–580. https://doi.org/10.1029/GL006i007p00577

Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2018). Multivariate lstm-fcns for time series classification. *arXiv preprint arXiv*, *1801*, 04503.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv*, *1412*, 6980.

Kullen, A., & Karlsson, T. (2004). On the relation between solar wind, pseudobreakups, and substorms. *Journal of Geophysical Research*, *109*, A12218. https://doi.org/10.1029/2004JA010488

Kullen, A., Ohtani, S., & Karlsson, T. (2009). Geomagnetic signatures of auroral substorms preceded by pseudobreakups. *Journal of Geophysical Research*, *114*, A04201. https://doi.org/10.1029/2008JA013712

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436. https://doi.org/10.1038/nature14539

Lee, W.-M. (2019). *Supervised Learning—Classification Using Logistic Regression* (Chap. 7, pp. 151–175). Indianapolis, IN: John Wiley and Sons, Ltd. https://doi.org/10.1002/9781119557500.ch7

Lee, D.-Y., Choi, K.-C., Ohtani, S., Lee, J. H., Kim, K. C., Park, K. S., & Kim, K.-H.(2010). Can intense substorms occur under northward imf conditions? *Journal of Geophysical Research*, *115*, A01211. https://doi.org/10.1029/2009JA014480

Li, X., Oh, K. S., & Temerin, M. (2007). Prediction of the al index using solar wind parameters. *Journal of Geophysical Research*, *112*, A06224. https://doi.org/10.1029/2006JA011918

Liou, K. (2010). Polar ultraviolet imager observation of auroral breakup. *Journal of Geophysical Research*, *115*, A12219. https://doi.org/10.1029/2010JA015578

Lui, A. T. Y. (1991). A synthesis of magnetospheric substorm models. *Journal of Geophysical Research*, *96*, 1849–1856. https://doi.org/10.1029/90JA02430

Lui, A. T. Y. (1996). Current disruption in the earth's magnetosphere: Observations and models. *Journal of Geophysical Research*, *101*, 13,067–13,088. https://doi.org/10.1029/96JA00079

Lui, A. T. Y. (2000). Tutorial on geomagnetic storms and substorms. *IEEE Transactions on Plasma Science*, *28*(6), 1854–1866. https://doi.org/10.1109/27.902214

Luo, B., Li, X., Temerin, M., & Liu, S. (2013). Prediction of the au, al, and ae indices using solar wind parameters. *Journal of Geophysical Research: Space Physics*, *118*, 7683–7694. https://doi.org/10.1002/2013JA019188

Lyons, L. R. (1995). A new theory for magnetospheric substorms. *Journal of Geophysical Research*, *100*, 19,069–19,081. https://doi.org/10.1029/95JA01344

Lyons, L. R., Blanchard, G. T., Samson, J. C., Lepping, R. P., Yamamoto, T., & Moretto, T.(1997). Coordinated observations demonstrating external substorm triggering. *Journal of Geophysical Research*, *102*, 27,039–27,051. https://doi.org/10.1029/97JA02639

McComas, D., Bame, S., Barker, P., Feldman, W., Phillips, J., Riley, P., & Griffee, J. (1998). Solar wind electron proton alpha monitor (swepam) for the advanced composition explorer. *Space Science Reviews*, *86*(1), 563–612. https://doi.org/10.1023/A:1005040232597

McKinney, W. (2010). Data structures for statistical computing in python. In W. McKinney (Ed.), *Proceedings of the 9th Python in Science Conference* pp. 51–56). Austin, TX: SciPy Conference.

McPherron, R. L. (1979). Magnetospheric substorms. *Reviews of Geophysics*, *17*, 657–681. https://doi.org/10.1029/RG017i004p00657

McPherron, R. L., & Chu, X. (2016). Relation of the auroral substorm to the substorm current wedge. *Geoscience Letters*, *3*(1), 12. https://doi.org/10.1186/s40562-016-0044-5

McPherron, R. L., Russell, C. T., & Aubry, M. P. (1973). Satellite studies of magnetospheric substorms on august 15, 1968: 9. phenomenological model for substorms. *Journal of Geophysical Research*, *78*(16), 3131–3149. https://doi.org/10.1029/JA078i016p03131

McPherron, R. L., Weygand, J. M., & Hsu, T.-S. (2008). Response of the earth's magnetosphere to changes in the solar wind. *Journal of Atmospheric and Solar-Terrestrial Physics*, *70*(2), 303–315. recent Observations and Simulations of the Sun-Earth System https://doi.org/10.1016/j.jastp.2007.08.040

Millman, K. J., & Aivazis, M. (2011). Python for scientists and engineers. *Computing in Science & Engineering*, *13*(2), 9–12. https://doi.org/10.1109/MCSE.2011.36

Miyashita, Y., Kamide, Y., Liou, K., Wu, C.-C., Ieda, A., Nishitani, N., et al. (2011). Successive substorm expansions during a period of prolonged northward interplanetary magnetic field. *Journal of Geophysical Research*, *116*, A09221. https://doi.org/10.1029/2011JA016719

Morley, S. K., & Freeman, M. P. (2007). On the association between northward turnings of the interplanetary magnetic field and substorm onsets. *Geophysical Research Letters*, *34*, L08104. https://doi.org/10.1029/2006GL028891

Newell, P., & Gjerloev, J. W. (2011a). Evaluation of supermag auroral electrojet indices as indicators of substorms and auroral power. *Journal of Geophysical Research*, *116*, A12211. https://doi.org/10.1029/2011JA016779

Newell, P., & Gjerloev, J. W. (2011b). Substorm and magnetosphere characteristic scales inferred from the supermag auroral electrojet indices. *Journal of Geophysical Research*, *116*, A12232. https://doi.org/10.1029/2011JA016936

Newell, P., Gjerloev, J. W., & Mitchell, E. J. (2013). Space climate implications from substorm frequency. *Journal of Geophysical Research: Space Physics*, *118*, 6254–6265. https://doi.org/10.1002/jgra.50597

Newell, P., & Liou, K. (2011). Solar wind driving and substorm triggering. *Journal of Geophysical Research*, *116*, A03229. https://doi.org/10.1029/2010JA016139

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. https://doi.org/10.1016/j.jastp.2016.04.019

Newell, P., Sotirelis, T., Liou, K., Meng, C. I., & Rich, F. J. (2007). A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *Journal of Geophysical Research*, *112*, A01206. https://doi.org/10.1029/2006JA012015

O'Brien, T. P. (2009). Seaes-geo: A spacecraft environmental anomalies expert system for geosynchronous orbit. *Space Weather*, *7*, 09003. https://doi.org/10.1029/2009SW000473

Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. https://doi.org/10.1109/MCSE.2007.53

Petrukovich, A. A., Baumjohann, W., Nakamura, R., Mukai, T., & Troshichev, O. A. (2000). Small substorms: Solar wind input and magnetotail dynamics. *Journal of Geophysical Research*, *105*, 21,109–21,117. https://doi.org/10.1029/2000JA900057

Pudovkin, M. (2013). Physics of Magnetospheric Substorms: A Review. In J. R. Kan, T. A. Potemra, S. Kokubun, & T. Iijima (Eds.), *Magnetospheric Substorms*. https://doi.org/10.1029/GM064p0017

Pulkkinen, T. I., Goodrich, C. C., & Lyon, J. G. (2007). Solar wind electric field driving of magnetospheric activity: Is it velocity or magnetic field? *Geophysical Research Letters*, *34*, L21101. https://doi.org/10.1029/2007GL031011

Ripley, B. D. (1996). *Pattern recognition and neural networks /*Cambridge: Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, C. T. (2000). How northward turnings of the imf can lead to substorm expansion onsets. *Geophysical Research Letters*, *27*, 3257–3259. https://doi.org/10.1029/2000GL011910

Russell, S. J., & Norvig, P. (2010). Artificial intelligence :, (3rd)., *Prentice Hall series in artificial intelligence*. Upper Saddle River, N.J.: Prentice Hall.

Sandholt, P. E., Farrugia, C. J., Moen, J., & Cowley, S. W. H. (1998). Dayside auroral configurations: Responses to southward and northward rotations of the interplanetary magnetic field. *Journal of Geophysical Research*, *103*, 20,279–20,295. https://doi.org/10.1029/98JA01541

Smith, C., L'Heureux, J., Ness, N., Acuña, M., Burlaga, L., & Scheifele, J. (1998). The ace magnetic fields experiment. *Space Science Reviews*, *86*(1), 613–632. https://doi.org/10.1023/A:1005092216668

Tanaka, T., Kitao, D., Sato, Y., Tanaka, Y., & Ikeda, D. (2015). Forecasting auroral substorms from observed data with a supervised learning algorithm. In *2015 IEEE 11th International Conference on e-Science*, pp. 284–287. Munich, Germany: https://doi.org/10.1109/eScience.2015.48

Tanskanen, E. I. (2009). A comprehensive high-throughput analysis of substorms observed by image magnetometer network: Years 1993-2003 examined. *Journal of Geophysical Research*, *114*, A05204. https://doi.org/10.1029/2008JA013682

Temerin, M., & Li, X. (2006). Dst model for 1995-2002. *Journal of Geophysical Research*, *111*, A04221. https://doi.org/10.1029/2005JA011257

Troshichev, O. A., Kotikov, A. L., Bolotinskaya, B. D., & Andrezen, V. G. (1986). Influence of the imf azimuthal component on magnetospheric substorm dynamics. *Journal of geomagnetism and geoelectricity*, *38*(11), 1075–1088. https://doi.org/10.5636/jgg.38.1075

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, IEEE, pp. 1578–1585.

Weigel, R. S., Klimas, A. J., & Vassiliadis, D. (2003). Solar wind coupling to and predictability of ground magnetic fields and their time derivatives. *Journal of Geophysical Research*, *108*(A7), 1298. https://doi.org/10.1029/2002JA009627

Wild, J. A., Woodfield, E. E., & Morley, S. K. (2009). On the triggering of auroral substorms by northward turnings of the interplanetary magnetic field. *Annales Geophysicae*, *27*(9), 3559–3570. https://doi.org/10.5194/angeo-27-3559-2009