Developing a Neural Network for the Identification of EMIC Wave Events

Nyssa S. S. Capman[1], Laura E. Simms[2], and Mark J. Engebretson[3]

[1]University of Minnesota
[2]Augsburg University
[3]Department of Physics, Augsburg University

September 11, 2023

Abstract

A supervised convolutional neural network (CNN) was developed to automatically identify electromagnetic ion cyclotron (EMIC) wave events from spectrograms. These events have usually been identified manually, which can be a time-consuming process. Statistical analyses of larger datasets would be facilitated if this process were simplified. The neural network model was trained on spectrogram images from the Halley magnetometer station that had been manually identified as either containing or not containing an EMIC wave event anywhere in the spectrogram. This model was tested on an unseen set of spectrograms, achieving a perfect true positive rate of 1. Size, time, frequency, and pixel color information was extracted from each identified event and exported into a spreadsheet for easier analysis. This method has the capability of reducing time and effort required to identify important spectrogram features by hand. Such an automated method could be applied to other space weather data stored in spectrograms.

Hosted file

972311_0_art_file_11336646_s07m4p.docx available at https://authorea.com/users/658668/articles/662921-developing-a-neural-network-for-the-identification-of-emic-wave-events

Developing a Neural Network for the Identification of EMIC Wave Events

Nyssa S. S. Capman[1], Laura E. Simms[2,3], Mark J. Engebretson[2]

[1]Department of Mechanical Engineering, University of Minnesota, Twin Cities, Minneapolis, MN, USA

[2]Department of Physics, Augsburg University, Minneapolis, MN, USA

[3]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA

Key Points

- Automatic identification of EMIC wave events in spectrograms is accomplished with an image classifying convolutional neural network
- Information about the size, time and frequency range, and power of each event is produced and output to a spreadsheet for analysis
- The method applied to 3 years of spectrograms from the Halley, Antarctica ground magnetometer station gives a true positive rate of 1

Abstract

A supervised convolutional neural network (CNN) was developed to automatically identify electromagnetic ion cyclotron (EMIC) wave events from spectrograms. These events have usually been identified manually, which can be a time-consuming process. Statistical analyses of larger datasets would be facilitated if this process were simplified. The neural network model was trained on spectrogram images from the Halley magnetometer station that had been manually identified as either containing or not containing an EMIC wave event anywhere in the spectrogram. This model was tested on an unseen set of spectrograms, achieving a perfect true positive rate of 1. Size, time, frequency, and pixel color information was extracted from each identified event and exported into a spreadsheet for easier analysis. This method has the capability of reducing time and effort required to identify important spectrogram features by hand. Such an automated method could be applied to other space weather data stored in spectrograms.

Plain language summary

Electromagnetic ion cyclotron waves in the earth's magnetosphere are often represented in spectrogram images, and wave events have usually been identified by a human examining the images by eye. To speed up the data collection process, an automatic method to extract information about each event has been developed.

1. Introduction

Space weather wave events have usually been identified by examining spectrograms by eye and recording the frequency and time ranges of notable wave events. This process is time-consuming and slows down the analysis of space weather data even as the amount of available data has increased. Previous work has sought to develop methods to automate this process, including use of Fourier Transform methods to analyze the power spectral density of wave data (Bortnik et al., 2007; Kim et al., 2018; Di Matteo et al., 2021; Inglis et al., 2015; Inglis et al., 2016; Murphy et al., 2020), discrete wavelet transforms (Omondi et al., 2022), and trigger algorithms that look for cases of simultaneity between two or more variables known to correlate with the events of interest (Carson et al., 2013). Few studies have explored the use of image analysis and object identification algorithms to automatically detect wave events directly from spectrogram images (Antonopoulou et al., 2022). However, neural network image analysis is a common method of object identification in remote sensing, used to locate and count craters (DeLatte et al., 2019), map water levels (Mandlburger et al., 2021), and map coral reefs (Li et al., 2020) among other applications. In all of these examples, the basic task is the same as is needed for space weather spectrograms: identifying shapes in image data and recording their location and characteristics such as size and color.

This work shows a convolutional neural network (CNN) developed to identify electromagnetic ion cyclotron (EMIC) wave events in spectrograms measured at the Halley, Antarctica ground magnetometer station between November 2006 and December 2009. The algorithm identifies and records both the time frame and frequency range of each EMIC wave event. The algorithm also outputs a rough estimate for wave power by recording the ratio of pixels in several color bins to the total number of pixels in the event. This aids the human researcher to filter which events should be further examined, or which events should be included in statistical analyses depending on particular research needs.

2. Model Training and Event Extraction

2.1. Model Training

A convolution neural network (CNN) was trained using the open-source PyTorch framework in Python (Paszke et al., 2019). A set of 1130 spectrograms from the Halley, Antarctica magnetometer station spanning November 2006 to December 2009 were used to train and test the model. These were manually divided into three classes: “no signals” (483 spectrograms), “broadband signals” (324 spectrograms), or “EMIC events” (323 spectrograms). The entire dataset was split into training and test sets in a 70:30 ratio. This was done using stratified sampling in order to maintain the same overall class distribution in each set. The CNN was trained on the training set, and then tested on the “unseen” spectrograms in the test set. The predicted classes were compared to the true classes and a confusion matrix was produced from which various metrics including the overall accuracy and the true positive rates of each class were calculated. Fifty epochs were used to train the models. This number was chosen by examining the training and validation loss, which did not show significant decreases after this point. The CNN consisted of a linear layer, a Tanh activation layer, a second linear layer, and finally a log softmax layer. A batch size of 64 was used. The Python code used to perform this model training has been uploaded to a Zenodo repository (Capman, 2023).

2.2. Event Extraction

After model testing, a Laplacian of Gaussian (LoG) "blob detection" algorithm was applied to only the spectrograms classified as containing events (Python library scikit-image, van der Walt et al., 2014). This algorithm produces circles indicating the location of identified events in pixels. A lookup table was created to relate the pixel locations with the time and frequency information on the x- and y-axes in the spectrograms. This lookup table depends on all spectrograms input into the algorithm having the same layout, dimensions, and axis limits. The identified events are automatically recorded in a spreadsheet which contains the approximate frequency and time frame of the events. This method determines the frequency and time frame of each event by recording the coordinates of the left-, right-, top-, and bottom-most edges of the circle produced by the LoG method. As the LoG identification circles do not always precisely encircle each event, these edges are only estimates. However, they do give a general location in each axis. Also recorded is the total number of pixels in each event circle and the number of pixels in each event circle belonging to each of six color bins (white, red, orange, yellow, green, and blue/black). These color bins were defined in relation to the colors used for plotting the wave power, and as such, the ratio of pixels in each color bin to total pixels in the event circle can be used as a rough estimate of the average power of each EMIC event. The total number of pixels in each event circle gives a rough estimate of the size of each EMIC event. These two pieces of information can help the user to determine which events are worth investigating.

3. Classification and Event Extraction Results

Three metrics are used to evaluate the success of the classification model: accuracy, true positive rate (TPR), and Heidke Skill Score (HSS). The accuracy is simply the ratio of correct predictions to total number of predictions. The TPR is calculated as TP/P, where TP is the number of true positives predicted and P is the total number of positives. Finally, the HSS is a measure of how well a model performed relative to the success of a random chance model (Heidke, 1926; calculations as in Ganushkina et al., 2015). The HSS can take values between negative infinity and +1, with +1 indicating a perfect prediction, 0 indicating that the model performed no better than random chance, and negative values indicating that the model performed worse than random chance. The model shown in the confusion matrix of Figure 1 has an accuracy of 86.4%, an HSS of 0.738, and most importantly a true positive rate of TPR = 1, when the "EMIC event" class is defined as the positive class, and both "no signals" and "broadband signals" as the negative classes. A TPR of 1 means that we correctly identified events 100% of the time. The accuracy and HSS weight false positives and false negatives equally. We consider the cost of false positives (i.e. a spectrogram with no event being identified as containing an event) to be much less than the cost of a false negative (i.e., missing an event). Therefore, the most important metric for this type of identification problem is the TPR.

The LoG algorithm identified larger events well. The algorithm also tended to identify many small and/or weak features that were not of interest, which were filtered out by a minimum radius cutoff. However, this cutoff requires some optimization to balance filtering out unimportant features with removing features of interest that happen to fall under the radius cutoff. For example, in Figure 2, Blobs 1 and 6 do not indicate EMIC events of interest and ideally would have been filtered out. However, increasing the minimum radius cutoff to exclude Blobs 1 and 6 would also have excluded Blob 4 which has the same radius as Blobs 1 and 6, but which contains a weak event of possible interest. This tradeoff is specific to each user's needs and the characteristics of each dataset.

It is important to note that broadband signals are only filtered out on a whole-spectrogram basis by the previous CNN step. For example, Blob 5 in Figure 2 is broadband but was still identified as a possible event. As this spectrogram had a strong event, it was classified into the "event" category despite the presence of broadband elsewhere in the spectrogram. The user must determine which identified blobs are useful and which are broadband based on size and overall color/wave power. Despite this drawback, the CNN/LoG combined method greatly reduces the work necessary to sort through spectrograms, and greatly limits the number of spectrograms to be examined by eye.

4. Discussion

One improvement needed for this method is in identifying true EMIC events in the presence of broadband signals in the blob detection algorithm. Currently, broadband is filtered out on a whole-spectrogram basis by the CNN classification. The model tends to classify spectrograms with large, clear events into the "EMIC event" class, regardless of the presence of broadband signal elsewhere in the spectrogram. Once these spectrograms are processed by the LoG blob detection algorithm, broadband signals are identified and recorded alongside the true EMIC events, and it is not always obvious from the spreadsheet output which events are broadband and which are true EMIC events. Conversely, the model classifies most spectrograms with strong broadband into the "broadband signal" class, regardless of the presence of smaller EMIC events, and as such these EMIC events are missed. Typically, broadband signals are identified by the LoG algorithm as being much larger than most EMIC events, since the broadband signals cover a larger portion of the spectrogram in the frequency axis. Based on this fact, the user may be able to filter out broadband signals by applying a maximum radius cutoff for each blob. However, this may also eliminate very large EMIC events which are especially important for analysis, and also not be capable of filtering out smaller broadband blobs, such as Blob 5 in Figure 2. Therefore, such a solution has a high potential cost. Further work is needed to improve the discrimination between true EMIC events and broadband signals.

In this study, we opted to divide the data into 3 classes ("no signals", "broadband signals", and "EMIC events"). This improved the model success, since otherwise broadband signals were often confused with EMIC events. Another choice is to divide the data into 2 classes: "no signals", and "EMIC events". In this case, the user would have the option to alter the prediction threshold, making it more or less likely for a given spectrogram to be classified into either class. The softmax layer outputs a predicted probability that a given spectrogram is of one or the other class. The default classification threshold choice is 0.5: spectrograms with probabilities greater than or equal to 0.5 are placed in the positive class (“EMIC event”), and those with probabilities less than 0.5 are placed in the negative class (“no signals”). However, if there is a class imbalance in the dataset, or if one or the other misclassification is more costly, a threshold other than 0.5 might be appropriate, weighting more heavily towards the positive or negative classes. Finding an optimal threshold may be aided with the use of a ROC curve. However, we ultimately determined that dividing this data into three classes produced better results than to use two classes, regardless of threshold optimization.

5. Conclusions

The CNN/LoG combined method described here greatly reduces the time and effort required to identify spectrogram features by hand and could also be applied to other space weather data stored in spectrograms. With appropriate optimization, the method could eventually be used to rapidly produce a dataset of event statistics from a large set of spectrograms with little to no input from the human user.

Acknowledgements

N.S.S.C. was supported by the National Science Foundation (NSF) through the Graduate Research Fellowship Program (GRFP) program. EMIC wave activity data were obtained from the induction coil magnetometer located at the Halley, Antarctica, British Antarctic Survey (BAS) ground station at L-shell 4.6. This work was supported by NSF Grant AGS-2013648.

Data Availability Statement

Spectrograms from the Halley, Antarctica ground magnetometer station are available at http://space.augsburg.edu/searchcoil/index.html. The Python code used to train and test the CNN models, as well as the Halley, Antarctica spectrograms used for training are available at https://doi.org/10.5281/zenodo.8280090.

References

Antonopoulou, A., Balasis, G., Papadimitriou, C., Boutsi, A. Z., Rontogiannis, A., Koutroumbas, K., ... & Giannakis, O. (2022). Convolutional Neural Networks for Automated ULF Wave Classification in Swarm Time Series. *Atmosphere*, 13(9), 1488. https://doi.org/10.3390/atmos13091488

Bortnik, J., Cutler, J. W., Dunson, C., & Bleier, T. E. (2007). An automatic wave detection algorithm applied to Pc1 pulsations. *Journal of Geophysical Research: Space Physics*, 112(A4). https://doi.org/10.1029/2006JA011900

Capman, N. S. S. (2023). Python Code to Train a Neural Network for the Identification of EMIC Wave Events in Spectrograms. (Version 1.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.8280090

Carson, B. R., Rodger, C. J., & Clilverd, M. A. (2013). POES satellite observations of EMIC-wave driven relativistic electron precipitation during 1998–2010. *Journal of Geophysical Research: Space Physics*, 118(1), 232-243. https://doi.org/10.1029/2012JA017998

DeLatte, D. M., Crites, S. T., Guttenberg, N., & Yairi, T. (2019). Automated crater detection algorithms from a machine learning perspective in the convolutional neural network era. *Advances in Space Research*, 64(8), 1615-1628.

Di Matteo, S., Viall, N. M., & Kepko, L. (2021). Power spectral density background estimate and signal detection via the multitaper method. *Journal of Geophysical Research: Space Physics*, 126(2). https://doi.org/10.1029/2020JA028748

Ganushkina, N. Y., Amariutei, O. A., Welling, D., & Heynderickx, D. (2015). Nowcast model for low-energy electrons in the inner magnetosphere. *Space Weather*, 13(1), 16-34.

Heidke, P. (1926). Measures of success and goodness of wind force forecasts by the gale-warning service. Geogr. Ann., 8, 301–349. doi: https://doi.org/10.1080/20014422.1926.11881138

Inglis, A. R., Ireland, J., Dennis, B. R., Hayes, L., & Gallagher, P. (2016). A large-scale search for evidence of quasi-periodic pulsations in solar flares. *The Astrophysical Journal*, 833(2), 284. http://dx.doi.org/10.3847/1538-4357/833/2/284

Inglis, A. R., Ireland, J., & Dominique, M. (2015). Quasi-periodic pulsations in solar and stellar flares: Re-evaluating their nature in the context of power-law flare Fourier spectra. *The Astrophysical Journal*, 798(2), 108. http://dx.doi.org/10.1088/0004-637X/798/2/108

Kim, H., Hwang, J., Park, J., Bortnik, J., & Lee, J. (2018). Global characteristics of electromagnetic ion cyclotron waves deduced from Swarm satellites. *Journal of Geophysical Research: Space Physics*, 123(2), 1325-1336. https://doi.org/10.1002/2017JA024888

Li, J., Knapp, D. E., Fabina, N. S., Kennedy, E. V., Larsen, K., Lyons, M. B., ... & Asner, G. P. (2020). A global coral reef probability map generated using convolutional neural networks. *Coral Reefs*, 39, 1805-1815. doi: https://doi.org/10.1007/s00338-020-02005-6

Mandlburger, G., Kölle, M., Nübel, H., & Soergel, U. (2021). BathyNet: A deep neural network for water depth mapping from multispectral aerial images. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 89(2), 71-89. doi: https://doi.org/10.1007/s41064-021-00142-3

Murphy, K. R., Inglis, A. R., Sibeck, D. G., Watt, C. E. J., & Rae, I. J. (2020). Inner magnetospheric ULF waves: The occurrence and distribution of broadband and discrete wave activity. *Journal of Geophysical Research: Space Physics*, 125(9), e2020JA027887. https://doi.org/10.1029/2020JA027887

Omondi, S., Yoshikawa, A., Zahra, W. K., Fathy, I., & Mahrous, A. (2022). Automatic detection of auroral Pc5 geomagnetic pulsation using machine learning approach guided with discrete wavelet transform. *Advances in Space Research*. https://doi.org/10.1016/j.asr.2022.06.063

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32. https://doi.org/10.48550/arXiv.1912.01703

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... & Yu, T. (2014). scikit-image: image processing in Python. PeerJ, 2, e453. https://doi.org/10.7717/peerj.453

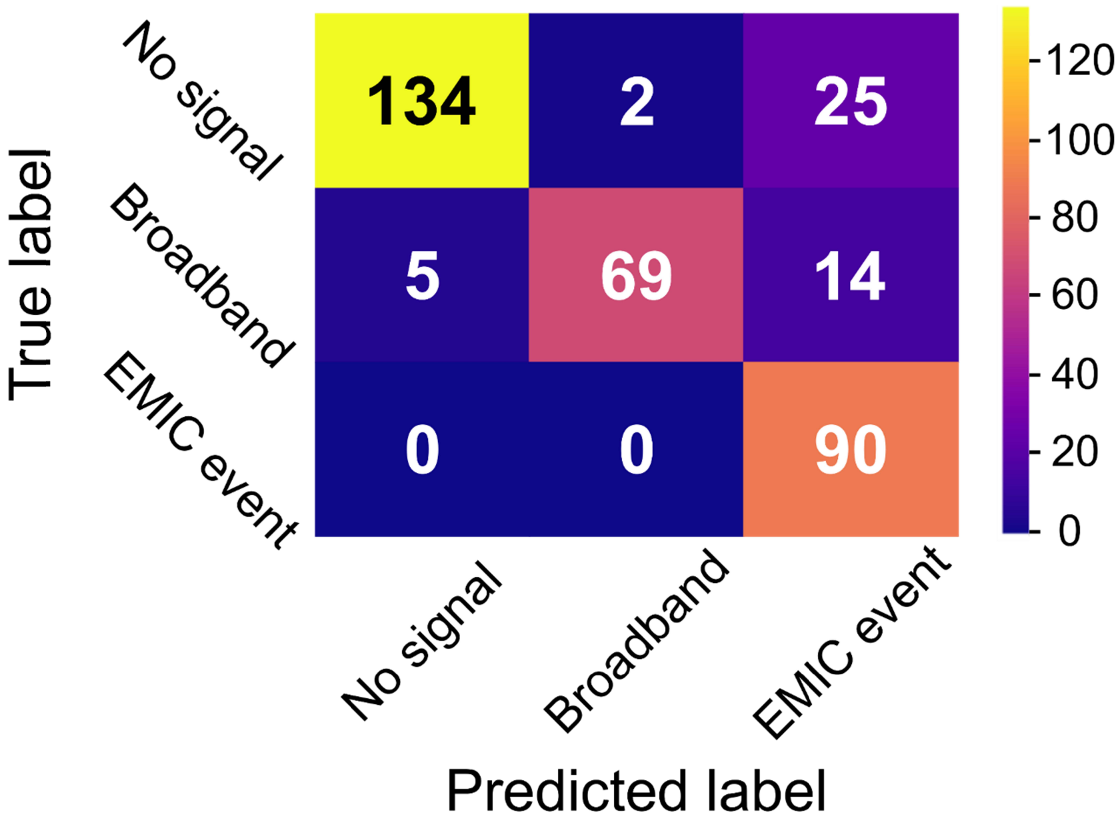

Figure 1. Confusion matrix for the CNN model showing true and predicted class labels for each spectrogram in the test set. Considering the “EMIC event” class as the positive class, the true positive rate (TPR) is 1. The overall accuracy is 86.4%, however, this includes misclassifications between the two negative, or non-event classes (“no signals” and “broadband signals”). The Heidke Skill Score (HSS) is 0.738.

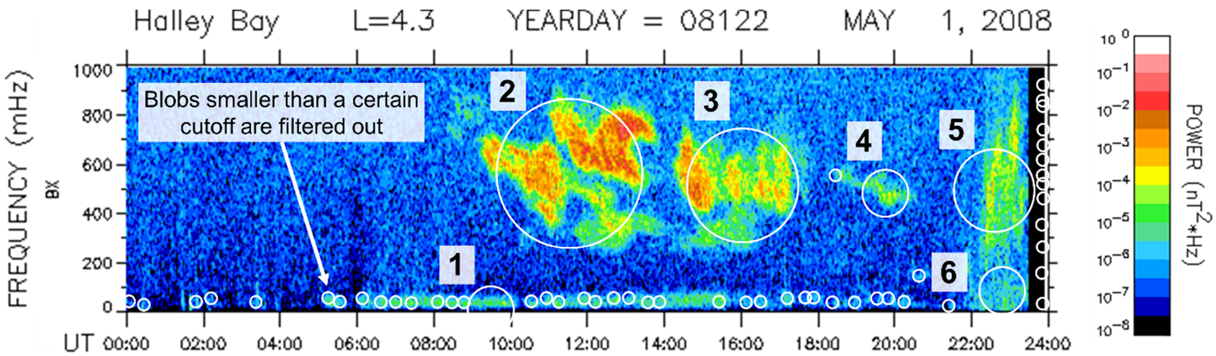

Figure 2. An example spectrogram indicating the LoG identification circles. The smaller, unnumbered circles would be filtered out before event information (frequency and time ranges, pixel counts in each color bin) is extracted and recorded in a spreadsheet. This filtering is based on a user-set minimum radius threshold. Blobs 1 and 6 were not filtered out by the radius cutoff set in this case, since filtering these out would have also eliminated Blob 4, a weak EMIC event. This tradeoff must be optimized by the user for the desired results and specific dataset being analyzed. Blobs 2 and 3 indicate strong EMIC events. Blob 5 contains broadband signal, and as such is a misidentification. Since broadband signals are filtered out on a whole-spectrogram basis by the CNN classification, some spectrograms containing broadband are classified into the “EMIC event” class since they also contain large, strong events (such as in Blobs 2 and 3).

Table 1. Spreadsheet output from the LoG identification of the spectrogram in Figure 2. For each blob identified, the approximate time and frequency ranges are recorded, as well as numbers of pixels in each color bin corresponding to different power ranges. These pixel color counts, along with the total number of pixels in the blob, help the user to make rough estimates of event size and average power. The user can then decide which events are worth investigating further or including in additional analyses.

			Pixel counts in each color bin						
Blob ID	**Approx. time range (hr)**	**Approx. freq. range (mHz)**	**# white 10^{-1} to 10^{0} nT^2Hz**	**# red 10^{-2} to 10^{-1} nT^2Hz**	**# orange 10^{-3} to 10^{-2} nT^2Hz**	**# yellow 10^{-4} to 10^{-3} nT^2Hz**	**# green 10^{-5} to 10^{-4} nT^2Hz**	**# blue, black 10^{-8} to 10^{-5} nT^2Hz**	**Total # in blob**
1	8-11	0-200	0	0	0	0	9	73	82
2	9-14	200-900	0	52	8	266	208	730	1264
3	14-18	200-800	0	3	5	172	200	386	766
4	19-21	400-600	0	0	0	1	16	116	133
5	21-24	300-700	0	0	0	22	119	299	440
6	22-24	0-200	0	0	0	0	4	133	137