URSI GASS 2023, Sapporo, Japan, 19 – 26 August 2023

Combining Unsupervised and Supervised Machine Learning for Lightning Classification: Application to Identifying EIPs for Ground-based TGF Detection

Yunjiao Pu[(1)], Steven A. Cummer*[(1)], Fanchao Lyu[(2)], Yu Zheng[(2)],
Michael S. Briggs[(3)], Stephen Lesage[(3)], Bagrat Mailyan[(4)], and Oliver J. Roberts[(5)]
(1) Duke University, Durham, NC, USA; e-mail: yunjiao.pu@duke.edu; cummer@duke.edu
(2) Nanjing Joint Institute for Atmospheric Sciences, Nanjing, China; e-mail: fanchao.lyu@gmail.com; zhengyu@cma.gov.cn
(3) University of Alabama in Huntsville, Huntsville, AL, USA; e-mail: briggsms@uah.edu; sjl0014@uah.edu
(4) Florida Institute of Technology, Melbourne, FL, USA; e-mail: mbagrat@gmail.com
(5) Universities Space Research Association, Huntsville, AL, USA; e-mail: oliver.roberts@nasa.gov

Abstract

We developed a machine learning framework to classify lightning radio signals and detect terrestrial gamma-ray flashes (TGFs). Energetic in-cloud pulses (EIPs, >150 kA) have been connected to a subset of TGFs, making it possible to detect TGFs on a continuous and large scale. However, manually searching for EIPs among many non-EIP events can be time-consuming, especially when using a lower peak current threshold. To address this issue, we used spectral clustering on low-dimensional features extracted from raw radio waveforms using an autoencoder. This revealed that +EIPs form a distinct class of waveform, comprising 6-7% of the total population. The resulting labeled dataset are used to train a supervised convolutional neural network (CNN) that targets +EIPs. Our CNN models identify on average 95.2% of true +EIPs with accuracy up to 98.7%, representing a powerful tool for +EIP classification. We then applied the pretrained CNN classifier to identify lower peak current EIPs (LEIPs, >50 kA) in a larger dataset. Among 10 LEIPs coincident with Fermi TGF observations, 2 previously reported TGFs and 2 unreported but suspected TGFs are found, while the majority were not associated with detectable TGFs, which suggests a more complicated LEIP-TGF relationship that calls for further study.

1 Introduction

Machine learning (ML) classifiers are a useful tool for automatically classifying images and waveforms from large datasets[1]. Previous studies have used ML to classify different types of lightning signals[2, 3], but the research and application of these approaches is still in its early stages. In this work, we developed a framework combining unsupervised clustering and supervised classification to explore patterns in a large dataset of lightning radio signals and classify complicated signals. We used this tool to identify energetic in-cloud pulses (EIPs, peak current >150 kA), which could be used to study terrestrial gamma-ray flashes (TGFs)[4, 5]. TGFs are energetic phenomena produced by relativistic runaway electron avalanche processes in thunderstorms[4], but they are difficult to observe. EIPs have been connected to a subset of TGFs, so studying EIPs can help improve our understanding of TGFs[6, 7]. We demonstrated the effectiveness of our ML-based tool for identifying EIPs, which could be used for large-scale, continuous detection of TGFs on the ground.

2 Data and Methodology

2.1 Dataset

We used ground-based low-frequency (LF) magnetic waveform data recorded near Duke University (DU) and Florida Institute of Technology (FT). The data was collected by sensors with two orthogonal magnetic coils that operated at approximately 1-300 kHz, with a sampling rate of 1 MS/s and time synchronized by GPS. We selected high peak current events based on peak current, polarity, time, and location provided by the U.S. National Lightning Detection Network (NLDN). From February 2020 to August 2021, we captured a total of 11,049 events that met these criterias. We also prepared a larger dataset with a lower positive NLDN polarity peak current threshold for identifying LEIP events. This dataset consisted of 32,775 events from March 2021 to December 2021. Meanwhile, we used gamma-ray data from the Fermi gamma-ray burst monitor (GBM) to verify if the detected LEIPs were coincident with TGF photon counts.

2.2 A machine learning framework for lightning classification

As shown in Fig1, a framework for classifying waveforms using a combination of unsupervised and supervised methods is developed. An unsupervised clustering model is used to explore patterns in a large dataset and group waveforms with similar features. These groups are then analyzed to determine their categories. The preliminary EIP/non-EIP labels are then manually refined and used to train a CNN model. The pretrained CNN model is then used to search for more EIPs in a larger dataset with a lower peak current

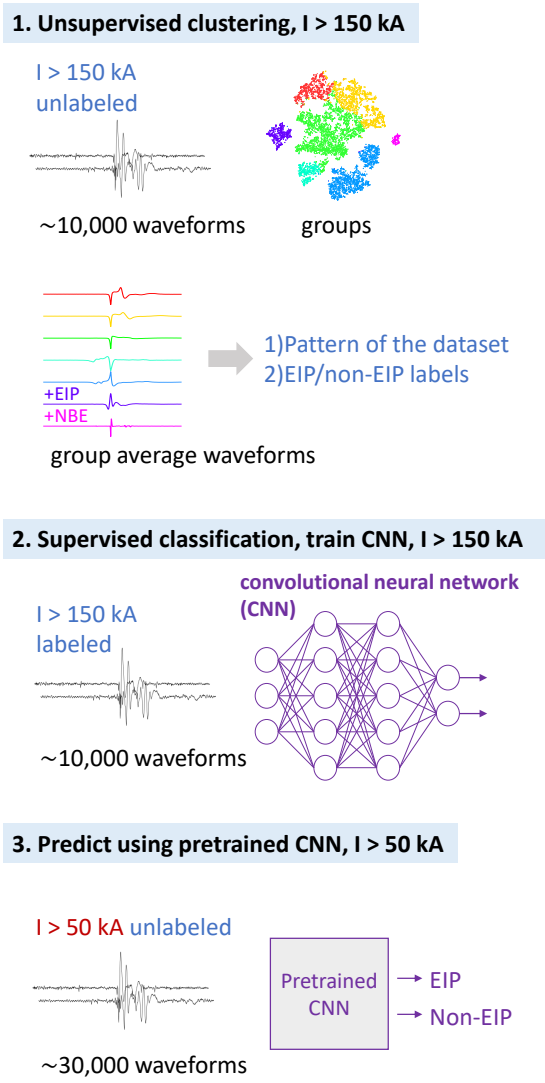


Figure 1. A framework merging unsupervised and supervised machine learning for lightning classification.

threshold between 50-150 kA, which was not feasible without ML.

2.3 Unsupervised clustering

Unsupervised clustering is used to understand the big lightning dataset. As illustrated in Fig.2, we performed spectral clustering on the features extracted by autoencoding. The autoencoder model consists of an encoder that encodes the input waveform into a small hidden layer and a decoder that reconstructs the original waveform from the hidden layer. The small hidden layer contains a compressed representation of the original waveform data, which is suitable for further spectral clustering. Analysis will be performed on the grouped clusters to understand the pattern of the entire dataset.

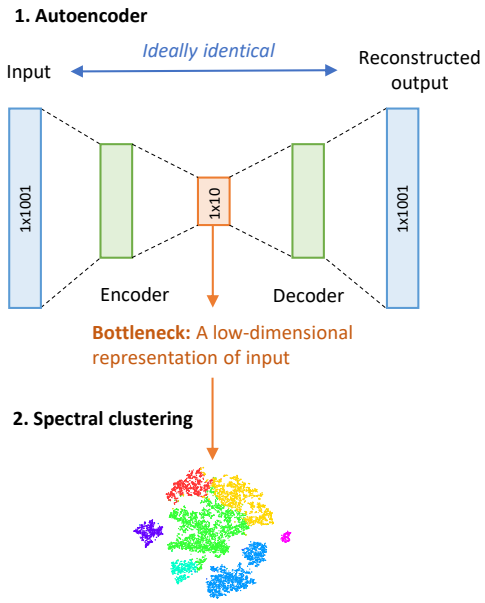


Figure 2. Illustration of unsupervised learning using an autoencoder model and spectral clustering.

2.4 Supervised classification with CNN

With a good understanding of the different kinds of energetic lightning events from the above unsupervised clustering as well as previous studies on lightning processes[5, 6, 7], we are able to pre-categorize these events and train a more effective classifier using supervised CNN classification. As shown in Fig.3, the CNN model consists of convolutional layers and classification layers, and uses the SoftMax function to normalize the output and classify the data into two categories: EIP and non-EIP. We used the "repeated K-folds validation" method to evaluate and tune the model's performance. In this work, we only adjust one hyperparameter, which is the weight of the EIP class in the loss function, in order to develop a model that has properly balanced EIP sensitivity and accuracy.

3 Results and Analysis

Fig.4b shows the results of clustering data using the t-SNE method. The spectral clustering algorithm divided the data into 7 groups. Fig.4c shows that the two most distinct groups are likely +EIPs and +NBEs, based on previous research[5]. Groups 1, 2, and 3 are thought to be +CGs with progressively smaller amplitudes of ionosphere-reflected sky waves. The nature of the lightning events in Groups 4 and 5 is less clear, so they are grouped together with Groups 1-3 as " +CGs and others".

Fig.4a presents the visualization and statistics of manually labeled events, which were divided into three categories: +EIPs (6.7%), +NBEs (1.9%), and +CGs and others (91.4%). These labels serve as a reference for evaluating the accuracy of the clustering model. The labels assigned to

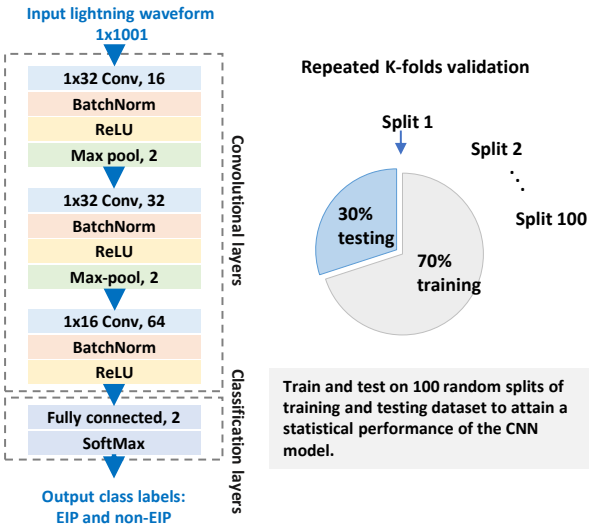


Figure 3. A framework merging unsupervised and supervised machine learning for lightning classification.

the grouped points by the spectral clustering algorithm for +EIPs correctly identified 90.7% of true +EIPs with an accuracy of 89.5%. However, many +EIPs were misclassified as belonging to Group 3 and Group 4 when the propagation distance was greater than approximately 600 km. Therefore, we decided to use a supervised classification model, which allows us to be more targeted in our EIP classification.

It is worth noting that +EIPs only make up 6-7% of the entire dataset, which means that the classification is being done on an imbalanced dataset with a majority of non-EIP events (∼93%) and a minority of EIP events. To address this issue, we trained four CNN models that assign different weights to the EIP and non-EIP classes in the loss function. The models, from CNN1 to CNN4, increasingly prioritize including as many true EIPs as possible, but this also affects the overall accuracy. We use standard definitions, with EIP sensitivity defined as the ratio of model-predicted true EIPs to all true EIPs, and EIP accuracy defined as the ratio of model-predicted true EIPs to all model-predicted EIPs. Our goal is to achieve both high EIP sensitivity and accuracy, but there is a tradeoff between the two measures.

Fig.5 illustrates the classification performance of the four CNN models. Each model was trained and tested on 100 random splits of the original dataset of >150 kA events. There are a total of 4×100=400 data points in Fig.5. The CNN classification models perform well with a mean EIP sensitivity of 95.2% and an EIP accuracy as high as 98.7%, CNN2 models appear to be best choice for practical use. This demonstrates that supervised CNN classification is a powerful approach for +EIP classification and for lightning classification in general.

We then applied the pretrained CNN2 model to search for +EIPs in a larger dataset of >50 kA lightning events, in or-

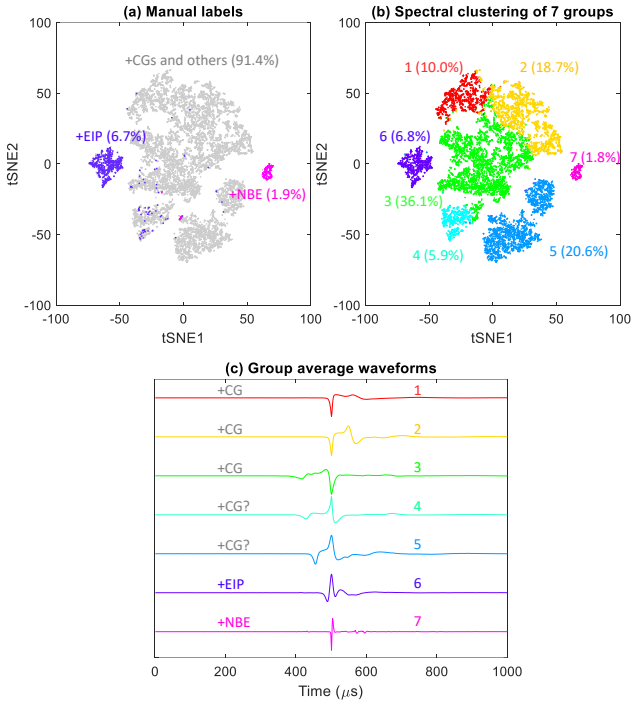
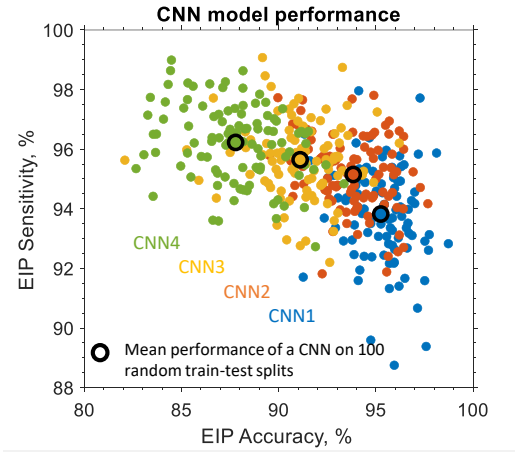


Figure 4. t-SNE visualization of clustering results and group average waveforms.



CNN1 Unweighted cross-entropy loss
CNN2 weighted cross-entropy loss, 0.05, 0.95
CNN3 weighted cross-entropy loss, 0.02, 0.98
CNN4 weighted cross-entropy loss, 0.01, 0.99

EIP sensitivity = predicted true EIPs / true EIPs
EIP accuracy = predicted true EIPs / predicted EIPs

Figure 5. Performance of 4 CNN models with different weights on EIP class in the loss function.

der to identify events with typical +EIP waveform shapes and potentially detect more TGFs that were previously not feasible without machine learning. Among the 32,775 events, the CNN model was highly confident (score >0.9) in classifying 998 events as +EIPs, which were considered as "true" +EIPs.

We examined the Fermi-GBM photon data at the time of these LEIPs to determine if they generated satellite-detectable TGFs. 10 LEIP events are found to be matched with Fermi within 600 km and 2 milliseconds. A case-by-case analysis of the LF waveforms and time-aligned photon counts of these 10 events revealed that 2 were definitive TGFs that had been previously reported by Fermi, and 2 were suspected TGFs that had not been reported by Fermi but had a small peak (verified to not be caused by cosmic rays) higher than the environmental noise in the photon profile. These 4 cases are shown in Fig.6. However, the remaining 6 LEIPs did not appear to be associated with detectable TGFs. It is unclear whether this is because the TGF source was located deep in the cloud and not bright enough to be detected by the satellite, or simply because some LEIPs do not produce TGFs. These results suggest that the relationship between TGFs and EIPs of lower peak current (particularly those in the 50-100 kA range) is complex and requires further study.

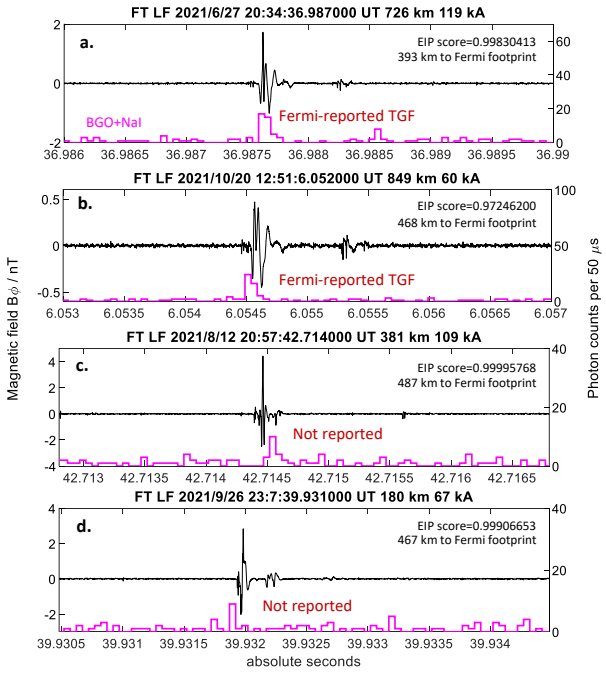


Figure 6. LF waveforms and Fermi photon counts for 4 +LEIPs. 2 are Fermi-reported TGFs and 2 are suspected TGFs but not reported by Fermi.

4 Summary

We developed machine learning classifiers that combine unsupervised and supervised methods to accurately identify +EIPs (>150 kA) with high sensitivity. These classifiers represent a powerful tool for monitoring EIP-type TGFs and classifying lightning events in general. When we applied these classifiers to lower peak current EIPs, we found that the majority of EIPs in the range of 50-100 kA did not seem to be associated with detectable TGFs. This finding warrants further investigation.

Acknowledgements

This study was supported by the National Science Foundation Dynamic and Physical Meteorology program through grant AGS-2026304. The authors would like to thank Amitabh Nag, Anjing Huang, Hamid Rassoul, Hamza Khounate, Mathieu Plaisir, and Naomi Watanabe for their assistance with the LF system, and Melissa Gibby for help with accessing Fermi data. They also thank Simiao Ren and Joe Lucas for helpful discussions about machine learning.

References

[1] Alzubaidi, L., J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan (2021), Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, *Journal of big Data*, 8(1), 1-74.

[2] Wang, J., Q. Huang, Q. Ma, S. Chang, J. He, H. Wang, X. Zhou, F. Xiao, and C. Gao (2020), Classification of VLF/LF Lightning Signals Using Sensors and Deep Learning Methods, *Sensors*, 20(4), 1030.

[3] Zhu, Y. A., P. Bitzer, V. Rakov, and Z. Q. Ding (2021), A Machine-Learning Approach to Classify Cloud-to-Ground and Intracloud Lightning, *Geophysical Research Letters*, 48(1), doi:10.1029/2020gl091148.

[4] Fishman, G. J., P. Bhat, R. Mallozzi, J. Horack, T. Koshut, C. Kouveliotou, G. Pendleton, C. Meegan, R. Wilson, and W. Paciesas (1994), Discovery of intense gamma-ray flashes of atmospheric origin, *Science*, 264(5163), 1313-1316.

[5] Lyu, F. C., S. A. Cummer, and L. McTague (2015), Insights into high peak current in-cloud lightning events during thunderstorms, *Geophysical Research Letters,* 42(16), 6836-6843, doi:10.1002/2015gl065047.

[6] Lyu, F. C., et al. (2016), Ground detection of terrestrial gamma ray flashes from distant radio signals, *Geophysical Research Letters*, 43(16), 8728-8734, doi:10.1002/2016gl070154.

[7] Lyu, F. C., S. A. Cummer, M. Briggs, D. M. Smith, B. Mailyan, and S. Lesage (2021), Terrestrial Gamma-Ray Flashes Can Be Detected With Radio Measurements of Energetic In-Cloud Pulses During Thunderstorms, *Geophysical Research Letters*, 48(11), doi:10.1029/2021gl093627.