Check for updates

JGR Atmospheres

RESEARCH ARTICLE
10.1029/2022JD038369

Key Points:

- A framework merging unsupervised clustering and supervised convolutional neural network (CNN) for lightning classification is developed
- Clustering of positive polarity energetic lightning radio pulses (>150 kA) identifies three processes: +EIPs (6%–7%), +NBEs, and +CGs
- CNNs detect 95.2% of manually identified +EIPs with up to 98.7% accuracy, enabling studying EIP-TGF link with lower peak current (>50 kA)

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. A. Cummer,
cummer@duke.edu

Citation:

Pu, Y., Cummer, S. A., Lyu, F., Zheng, Y., Briggs, M. S., Lesage, S., et al. (2023). Unsupervised clustering and supervised machine learning for lightning classification: Application to identifying EIPs for ground-based TGF detection. *Journal of Geophysical Research: Atmospheres*, *128*, e2022JD038369. https://doi.org/10.1029/2022JD038369

Received 22 DEC 2022
Accepted 18 APR 2023

© 2023. American Geophysical Union. All Rights Reserved.

Unsupervised Clustering 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[2,3], Yu Zheng[2], Michael S. Briggs[4], Stephen Lesage[5], Bagrat Mailyan[6], and Oliver J. Roberts[7]

[1]Electrical and Computer Engineering Department, Duke University, Durham, NC, USA, [2]Key Laboratory of Transportation Meteorology of China Meteorological Administration, Nanjing Joint Institute for Atmospheric Sciences, Nanjing, China, [3]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China, [4]Center for Space Plasma and Aeronomic Research, University of Alabama in Huntsville, Huntsville, AL, USA, [5]Department of Space Sciences, University of Alabama in Huntsville, Huntsville, AL, USA, [6]Florida Institute of Technology, Melbourne, FL, USA, [7]Universities Space Research Association (USRA), Huntsville, AL, USA

Abstract We developed a framework merging unsupervised and supervised machine learning to classify lightning radio signals, and applied it to the possible detection of terrestrial gamma-ray flashes (TGFs). Recent studies have established a tight connection between energetic in-cloud pulses (EIPs, >150 kA) and a subset of TGFs, enabling continuous and large-scale ground-based TGF detection. However, even with a high peak current threshold, it is time-consuming to manually search for EIPs in a background of many non-EIP events, and it becomes even more difficult when a lower peak current threshold is used. Machine learning classifiers are an effective tool. Beginning with unsupervised learning, spectral clustering is performed on the low-dimensional features extracted by an autoencoder from raw radio waveforms, showing that +EIPs naturally constitute a distinct class of waveform and 6%–7% of the total population. The clustering results are used to form a labeled data set (~10,000 events) to further train supervised convolutional neural network (CNN) that targets for +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. The pretrained CNN classifier is further applied to identify lower peak current EIPs (LEIPs, >50 kA) from a larger data set (~30,000 events). Among 10 LEIPs coincident with Fermi TGF observations, 2 previously reported TGFs and 2 unreported but suspected TGFs are found, while the majority are not associated with detectable TGFs. In addition, unsupervised clustering is found to reflect characteristics of the ionosphere reflection height and its effect on radio wave propagation.

Plain Language Summary In this study, we developed a machine learning-based method to classify lightning radio signals. The method uses unsupervised and supervised machine learning to distinguish different types of signals with high accuracy. The focus of the study is to identify energetic in-cloud pulses (EIPs) that are associated with a subset of lightning-related terrestrial gamma-ray flashes (TGFs). The method successfully identified 95.2% of true EIPs with up to 98.7% accuracy, and discovered new TGF events that were not previously reported. Additionally, the method revealed insights into the ionosphere and radio wave propagation. This method can be useful for studying lightning and other related phenomena.

1. Introduction

Machine learning (ML) classifiers are an effective tool to automatically classify images and waveforms from vast amount of data (Alzubaidi et al., 2021). Some previous studies have shown that ML approaches are capable of discriminating different lightning radio signals. With supervised learning, (J Wang et al., 2020) demonstrated how to classify 10 different types of lightning pulses using one-dimensional convolutional neural network (CNN), though their classification was performed upon selected waveforms that have a balanced number of 5,000 events for each class/type. Furthermore, (Zhu et al., 2021) applied Support Vector Machines approach to a realistic lightning data set to improve the classification accuracy of the In-cloud/Cloud-to-ground (IC/CG) lightning to 97%, better than the previous 91%−96% with multiparameter classification algorithms (Kohlmann et al., 2017). Nevertheless, the research and application of ML-based lightning classification approaches remains in its infancy.

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In this work, we developed a comprehensive framework that combines unsupervised clustering and supervised classification approaches to explore the underlying pattern of a big lightning radio data set and to effectively classify complicated lightning signals. We demonstrated the power of this ML-based tool with an example in identifying energetic in-cloud pulses (EIPs) (Lyu et al., 2015), which could benefit to the study of Terrestrial gamma-ray flashes (TGFs) (Lyu et al., 2016, 2021).

TGFs, as discovered by the Burst and Transient Source Experiment (BATSE) in 1991 (Fishman et al., 1994), are one of the most energetic phenomena that occur in the Earth's atmosphere. A prevalent theory is that TGFs are produced by relativistic runaway electron avalanche (RREA) processes in high-electric fields inside thunderstorms, especially during the upward in-cloud leader processes (Dwyer, 2003; Lu et al., 2010; Stanley et al., 2006). But there are significant ambiguities in the production of seed electrons and the location of high-electric fields where subsequent RREA processes occur, leading to different TGF models (Babich et al., 2015; Celestin & Pasko, 2011; Dwyer, 2012). The understanding of TGFs is largely limited by observations. TGFs are usually observed by orbital gamma-ray detectors, such as RHESSI (Grefenstette et al., 2009), AGILE (Marisaldi et al., 2010), Fermi-GBM (Briggs et al., 2013), BeppoSAX (Ursi et al., 2017), ASIM (Neubert et al., 2019) and Insight-HXMT (Zhang et al., 2021). Nevertheless, it is very difficult to capture TGFs since they occur randomly in thunderstorms at a rate way lower than lightning flashes and the gamma-ray detectors usually cover a small effective area of detection. For instance, Fermi-GBM detects on average ∼800 TGFs per year within 800 km of its spacecraft's nadir using BGO detectors of an effective area of ∼161 cm^2 (Roberts et al., 2018). Moreover, ground-based direct detection of TGFs could be even harder as a result of both the small likelihood of a downward TGF from natural lightning falling into the effective detection region of the detector, and the strong atmosphere attenuation near the ground, with only a handful of cases being reported (e.g., Dwyer et al., 2012; Tran et al., 2015; Wada et al., 2019).

In the meanwhile, recent studies have established a high-to-certain connection between EIPs (>150 kA) and a subset (∼10%) of TGFs (Lyu et al., 2015, 2016, 2021). These so-called EIP events are generated during the upward propagation of initial negative leaders and involve large peak currents above 150 kA. EIPs and TGFs have been proved to be two facets of a same phenomenon (Lyu et al., 2016), with the large-amplitude and relatively long-time scale (∼50 μs) EIP sferics possibly produced by the current from the TGF itself (Tilles et al., 2020) or a combination of currents from a hot channel and the TGF (Ostgaard et al., 2021). This opens a door for indirect observation of TGFs by ground radio measurements. Since the ground radio sensor usually continuously monitors a large area within hundreds to thousands of kilometers, such EIP-TGF connection enables unprecedentedly continuous and large-scale ground-based detection of TGFs (Lyu et al., 2021). Therefore, we can expect to establish a large data set of TGF-associated EIPs to investigate in detail the occurrence context of TGFs, overcoming the current limitations in time and space coverage in direct TGF observations. However, a new problem arises that even with a high peak current threshold (>150 kA), it is time-consuming to manually search for EIPs in a background of many non-EIP events, and it becomes even more difficult when a lower peak current threshold is used (e.g., >50 kA).

Machine learning classifiers are thus a perfect tool for this kind of big data classification problem. In this study, we combine unsupervised and supervised machine learning approaches to classify positive-polarity high peak current lightning signals and to identify +EIPs especially. There are a couple of technical questions that we are interested in: Is ML capable of identifying EIP-like events, which could be relatively rare, from a large number of data? How well does ML work comparing to humans? Can ML approaches identify EIPs without human knowledge? In other words, is unsupervised clustering capable of grouping EIPs spontaneously? Moreover, we also aim to investigate the following scientific problems: How many types of high peak current lightning pulses are there hidden in the lightning radio data? Are there EIP-like events that have lower peak currents below 150 kA? How are these lower peak current EIPs (LEIPs) related to TGFs?

As a result, our unsupervised clustering of high peak current lightning radio pulses (>150 kA) identifies three processes: +EIPs (6%−7%), positive narrow bipolar events (+NBEs, ∼2%), and +CGs (∼91%). Supervised CNN classifiers are trained upon the labeled clustering results and classify on average 95.2% of true +EIPs with accuracy up to 98.7%. Lower peak current +EIPs (LEIPs, 50−150 kA) are further found by the pretrained CNN and compared with Fermi TGF observations. Among the 10 LEIP-Fermi cases, 2 are Fermi-reported TGFs and 2 are unreported but suspected TGFs, while the majority (6 cases) are apparently not associated with detectable TGFs, suggesting a complicated LEIP-TGF connection. In addition, we also found that unsupervised clustering could automatically show the diurnal change of the ionosphere reflection height and its effect on radio wave propagation.

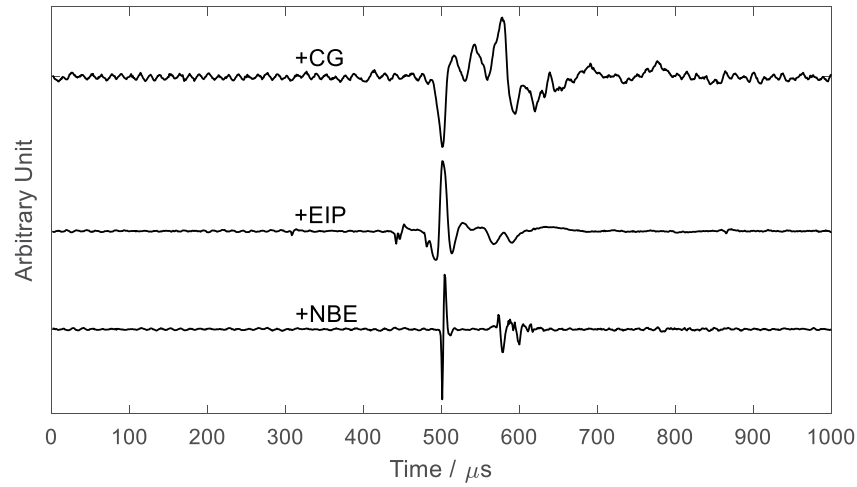

Figure 1. Example waveforms (1 × 1,001 points) of high peak current events (>150 kA) from the return stroke of a positive cloud-to-ground flash (+CG), positive energetic in-cloud pulse (+EIP), and positive narrow bipolar event (+NBE).

2. Data and Methodology

2.1. Data Preparation

This work uses ground-based LF magnetic waveform data recorded near Duke University (DU, 35.971°N, −97.440°E) and Florida Institute of Technology (FT, 28.062°N, −80.624°E). The LF sensors are composed of two orthogonal magnetic coils working at approximately 1–300 kHz. The sampling rate is 1 MS/s with time synchronized by GPS. The high peak current events are selected according to the peak current, polarity, time, and location provided by the U.S. National Lightning Detection Network (NLDN). The following two parameters are applied during the initial data preparation: distance to either of the two sensors ≤1,000 km and positive NLDN polarity peak current $I_{peak} \geq +150$ kA.

From February 2020 to August 2021, a total of 11,049 events that meet these criteria, including 4,719 FT events and 6,330 DU events, were captured by our remote sensors with magnetic waveforms stored for further classification. Note that a single lightning event might be detected by both DU and FT sensors, but for classification purposes, it would be treated as two independent events considering that the waveforms are not identical as a result of propagation differences. These event waveforms are normalized and centered at the peak absolute value in a time window from 0 to 1,000 μs (1 × 1,001 points). Example pulse waveforms of +CG, +EIP and +NBE events are shown in Figure 1. Note that we use the notation "+CG" to represent "+CG return stroke," as to maintain consistency with Lyu et al. (2015). The labeling of these events is semi-automated by doing quality control of the clustering results from unsupervised algorithms (see Section 2.2).

A second and larger data set formed with a lower positive NLDN polarity peak current threshold of 50 kA < I_{peak} < 150 kA but otherwise the same parameters is also prepared for +EIP classification. The data set consists of 32,775 events from FT from March 2021 to December 2021 during the time windows when the Fermi satellite is above the Americas, which accounts for ∼10% of the time of a day. This criterion is to limit the volume of radio data stored, and to enable a comparison with Fermi TGF observations. The data set will be used as a test data set for the search of lower peak current EIP-like events using the pre-trained CNN classifier.

In addition, gamma-ray data detected by Fermi gamma-ray burst monitor (GBM) are employed to verify if the ground-detected +EIPs are coincident with TGF photon counts. GBM consists of 2 bismuth germanate (BGO; ∼200 keV − 40 MeV) detectors and 12 sodium iodide (NaI; ∼8 keV − 1 MeV) detectors (Briggs et al., 2013). Both types of scintillator detectors manifest TGF production, but it should be noted that NaI detectors have uncertain arrival time delays due to Compton scattering (Østgaard et al., 2008) and longer instrumental deadtime (Briggs et al., 2013). The temporal relationship between +EIPs at LF and TGF gamma-ray counts will be investigated by aligning the two signals at the source position (Lyu et al., 2016, 2021; Pu et al., 2019).

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

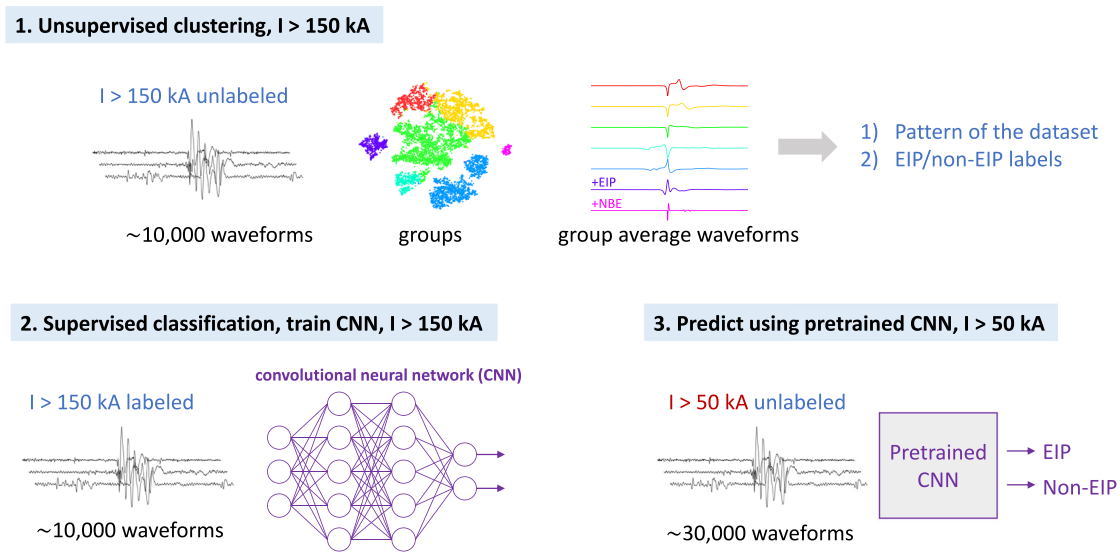

Figure 2. Framework of unsupervised and supervised classification of energetic radio pulses. Unsupervised clustering helps to visualize the pattern of the energetic radio waveforms (>150 kA) and provide initial labels. Supervised classification using convolutional neural networks (CNN) aims to train a specific model to identify EIPs with high performance. The pretrained CNN model will be applied to a larger data set to identify more EIP-like events but with lower peak currents (>50 kA).

2.2. Framework of Waveform Classification With Deep Learning

This work develops a comprehensive framework for waveform classification that merges unsupervised and supervised classification methods. As shown in Figure 2, an unsupervised clustering model is first implemented to explore the underlying pattern of a large data set. The basic idea is to cluster lightning waveforms into groups with similar features that are separated automatically by ML without interference of human knowledge. This step does not determine what the different groups are, but only determines that each group is different from the others. In the next step, these initial groups will be analyzed using supervised classification to determine their categories. For the purposes of this work, we focus only on distinguishing EIPs from non-EIPs.

While an unsupervised model by itself is capable of discriminating waveforms, a supervised model normally performs much better with the knowledge of true data. Thus, the preliminary >150 kA EIP/non-EIP labels are manually refined and then used to train a convolutional neural network (CNN). In this step, the goal is to train a model/classifier that can identify as many true +EIPs as possible while keeping low false alarms.

Finally, the pretrained CNN model will be used to search for +EIPs from a new data set that includes events of lower peak currents between 50 and 150 kA. The lower peak current data set includes too many events to classify manually. This step is to identify events with waveforms most similar to (of high likelihood) typical +EIPs shape but of relatively low amplitude, which has implications for many more but perhaps weak TGFs. A further comparison with Fermi GBM photon data will be made to investigate if these ground-detected +LEIPs are capable of generating satellite-detectable TGFs.

2.3. Unsupervised Clustering With Autoencoders and Spectral Clustering

The data set of ~10,000 > 150 kA lightning waveforms raises questions of what patterns are hidden in these high peak current events and how to visualize the similarity. Instead of relying on human assumptions and definitions, unsupervised clustering is applied to categorize similar waveforms impartially by measures of similarity/distance that are intrinsic to data itself. There are many types of clustering algorithms (Abbas, 2008; Gulati & Singh, 2015), such as K-means, spectral clustering, DBSCAN, Gaussian Mixture, etc. In this study, the way lightning data is organized in a graph-like structure called feature space can be complex. To handle this complexity, a clustering method called spectral clustering was used (Von Luxburg, 2007). This method focuses on how data points are related to each other in this feature space, instead of how close they are to each other (Singh, 2010).

Because spectral clustering is costly to compute for a large data set, we need to reduce the data dimension first by extracting the most useful features from raw lightning waveforms. In this work we apply autoencoding as the

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

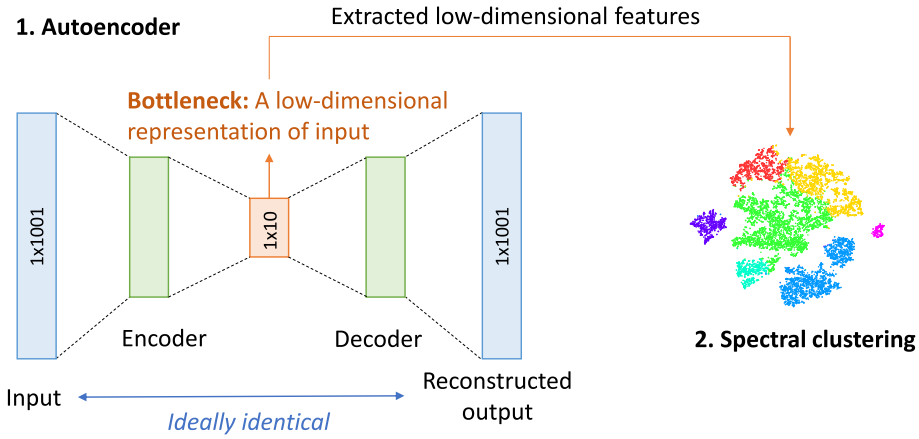

Figure 3. Illustration of spectral clustering using extracted low-dimensional features by an autoencoder model.

algorithm to compute a low-dimensional representation of the input data (Tschannen et al., 2018). As illustrated in Figure 3, the autoencoder model consists of an encoder that encodes the 1 × 1,001 input lightning waveform into a small-size hidden layer (1 × 10), and a decoder that reconstructs the original waveform from the hidden layer. After some training to minimize the difference between the input and the reconstructed output, the "bottleneck" 1 × 10 hidden layer contains almost all the information that needs to reconstruct the input and thus can be used as a compressed representation of the original large 1 × 1,001 waveform data for further spectral clustering (examples of original and reconstructed waveforms can be found in Figure S1 in Supporting Information S1). Analysis will be performed on the grouped clusters based on the 10-dimensional data in the bottleneck layer to recognize and understand the pattern of the entire data set.

2.4. Supervised Classification With Convolutional Neural Networks (CNNs)

The technical goal of this work is to automatically identify +EIPs with high sensitivity and accuracy from a large number of lightning LF waveforms. 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, we are able to pre-categorize these events and train a more effective classifier using supervised learning algorithms.

Different from clustering, grouping data points that are pre-categorized or labeled is known as classification. There are a lot of classical supervised classification algorithms in machine learning, such as logistic regression, K-nearest neighbors, decision trees, support vector machines, etc. But considering that lightning waveforms are complicated and unstructured (vs. structured data with well-defined features, e.g., "Weather" data that contain entries like "wind," "temperature," "'pressure," and so on), we will use deep learning to gather information that is not explicitly defined. The convolutional neural network (CNN) is one of the most useful deep learning algorithms to make classification on unstructured data of time series, images and videos (O'Shea & Nash, 2015).

The CNN model in this work is shown in Figure 4 and consists of convolutional and classification layers. The convolutional layers have three stacks with building blocks of Conv, BatchNorm, ReLU, and Max pool layers (O'Shea & Nash, 2015). The fully connected layer connects to a SoftMax layer for classification of two classes. The model uses forward propagation to extract information from raw lightning waveforms and downsample it to a low dimensional representation. Backpropagation is used in training to optimize weights and minimize loss (Alzubaidi et al., 2021). While there are many other architecture designs that may work, this CNN model is one of the most compact.

When training a CNN model, we apply the "repeated K-folds validation" method (Refaeilzadeh et al., 2009) to evaluate the performance and to tune the model hyperparameters. The energetic lightning data set that contains 11,049 events is randomly split into training and testing parts by a ratio of 70% versus 30%. For each split, the model is trained and tested once with the same model hyperparameters. We then repeat this process for 100 times and evaluate the statistical performance of the model. If the statistical performance of a CNN model is not satisfactory, we will move on to adjust hyperparameters of the CNN model which will then be regarded as a different CNN model. 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.

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

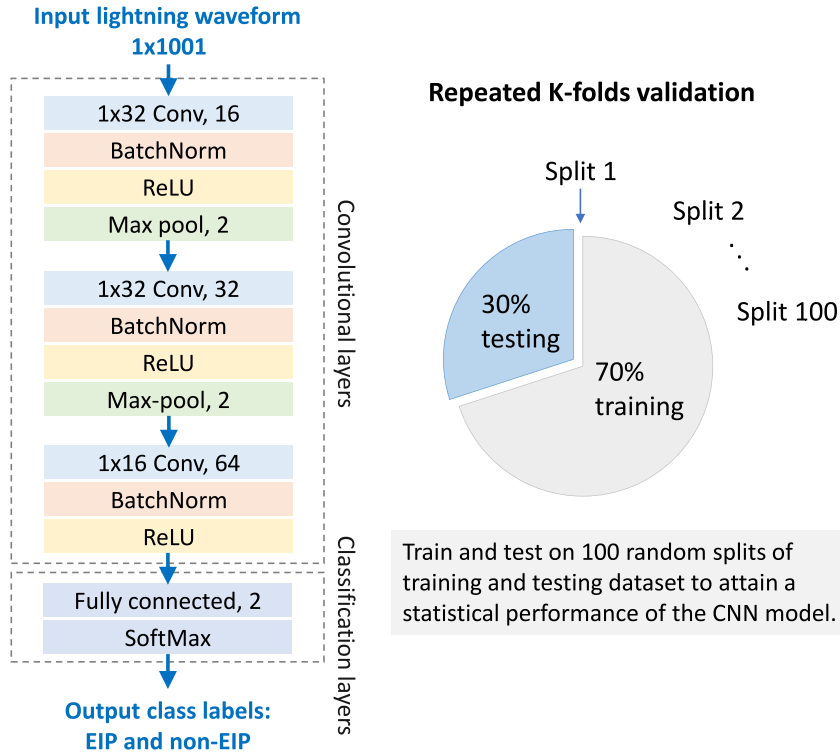

Figure 4. Architecture of the 1-dimensional convolutional neural network (CNN) and illustration of 100 random train/test splits at a ratio of 70/30. The CNN model is trained and tested 100 times to gain a statistical performance immune to the uncertainty due to train-test splits.

3. Results and Analysis

3.1. Unsupervised Clustering of High Peak Current Events >150 kA

Figure 5 shows clustering results of the 11,049 energetic lightning events. t-SNE method (Van der Maaten & Hinton, 2008) is applied here to reduce the 10-dimensional features identified by the autoencoder to 2-D and visualize the data. We let the spectral clustering algorithm cluster the data into seven groups, and the two most distinct groups are determined to be +EIPs and +NBEs. It's worth noting that the number of clusters should be larger than three known types and large enough to distinguish the EIP class, but not too many subgroups; seven

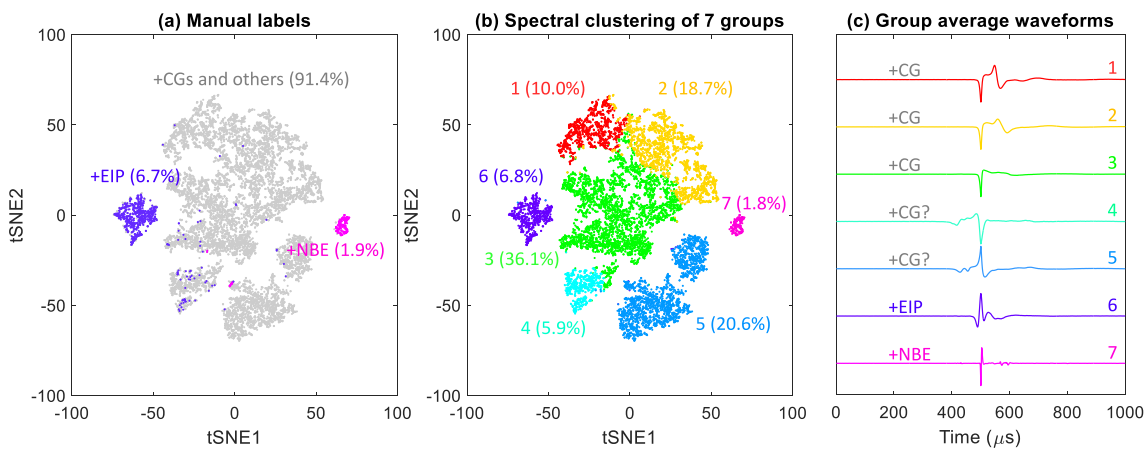

Figure 5. t-SNE visualization of 11,049 positive polarity high peak current events (>150 kA) with color coded by (a) manual labels, and (b) spectral clustering groups. (c) Average waveforms for each group from spectral clustering show that Group 6 and Group 7 are dominantly +EIPs and +NBEs, respectively, consistent with the manual labels.

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
Number of Positive Polarity High Peak Current (>150 kA) Events With Manual Class Labels

Manual labels	Number of events	Percentage %
+EIPs	743	6.7
+NBEs	214	1.9
+CGs and others	10,092	91.4
Total	11,049	100

clusters were deemed appropriate but similar numbers are also acceptable. Before we proceed to manually analyze the waveforms in each group, the group average waveforms are plotted out to show the prevalent features in each group as seen in Figure 5c. Compared with the example waveforms shown in Figure 1 and in previous studies about +EIPs (Lyu et al., 2015) and +NBEs (Eack, 2004), it is likely that Group 6 is dominated by +EIPs and Group 7 +NBEs.

Groups 1, 2, and 3 are likely +CGs with decreasing amplitude of ionosphere-reflected sky waves (the second pulse after 500 μs). This indicates that these energetic lightning events can be discriminated by their sky wave features and it worths a further look at how these subgroups of +CGs reflect the information of ionosphere (see Section 3.4). However, it is less clear what the lightning events are in Groups 4 and 5. A question may raise that are there any new categories of energetic lightning events that have not yet been recognized? For the time being, we apply question marks on these two unclear Groups 4 and 5, and we place all events from Groups 1 to 5 in a large group of "+CGs and others" for the sake of simplicity. More investigation on the above questions will be discussed in Section 3.4.

With this general understanding of the pattern of the data set, it becomes more straightforward to quality control the labels, especially for +EIPs which are the focus of this work. It is important to note that while the 3D parameter space defined by Lyu et al. (2015) was useful for identifying EIPs, it was not sufficient for dealing with larger datasets with more variations in individual EIPs and CGs. Unsupervised clustering as an independent classifier became helpful in identifying subtle features that could not be described by defined parameters. However, the ML model could generate visually apparent errors, and human reviewers had additional knowledge that the two waveforms from FT and DU sensors must have the same class label for a given EIP. Therefore, a hybrid approach combining human expertise and machine learning was used for manual labeling. This approach allowed human reviewers to ensure the most accurate representation of the EIPs by performing quality control on the results (examples of misclustered events are shown in Figure S2 in Supporting Information S1). It is worth noting that human reviewers depend on the typical waveforms of each type displayed in Figure 1 to identify the types of energetic lightning events accurately. However, for waveforms that appear ambiguous and lack simultaneous measurements from another station, both human and machine classifiers may misclassify them. To ensure consistency, two of our coauthors independently visually inspected all 11,049 events.

Figure 5a and Table 1 respectively show the visualization and statistics of the manually labeled events that are classified into three categories: +EIPs (6.7%), +NBEs (1.9%), and +CGs and others (91.4%). This manual labeling is assumed to be correct and thus the ground truth. The labels assigned to the grouped points in spectral clustering for +EIPs identifies 90.7% of the true EIPs with an accuracy of 89.5%, which is accurate enough to be useful. This also confirms that +EIPs and +NBEs are naturally two distinct classes of energetic lightning events.

Meanwhile, we notice that there are many +EIPs being misclassified into Group 3 and Group 4. Why are these +EIPs being misclassified by machine learning? It is found that the mean distance of ML-misidentified but true +EIPs is 772 km, while the mean distance of ML-identified true +EIPs is 629 km, suggesting that the ML clustering becomes relatively less sensitive to +EIPs when the propagation distance is above ~600 km. At longer distances, the event waveform gets more distorted by sky wave, thus it becomes more and more difficult to discriminate CGs and EIPs even by the human eye. This also indicates that the manually refined data set may still be subject to human error and is not necessarily exhaustive, particularly in cases where the observation distance is farther than 600 km. In this sense, ML performs similarly to human beings.

At this point, we could continue to use the unsupervised clustering model to discriminate +EIPs, and try to further improve the performance (Xie et al., 2016). Or we can switch to supervised classification using the labels/knowledge gained from unsupervised learning (clustering). We choose the latter, with the advantage that in supervised classification there is more flexibility in training the model to target for a specific known class.

3.2. Supervised Classification of +EIPs > 150 kA With CNN

Note that +EIPs are only 6%–7% of the entire data set, which means that the classification is made upon an imbalanced data set of a major non-EIP class (~93%) and a minor EIP class. Some special treatment needs to be made to balance the two classes (Tanha et al., 2020). Therefore, we trained 4 CNN models that put different

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

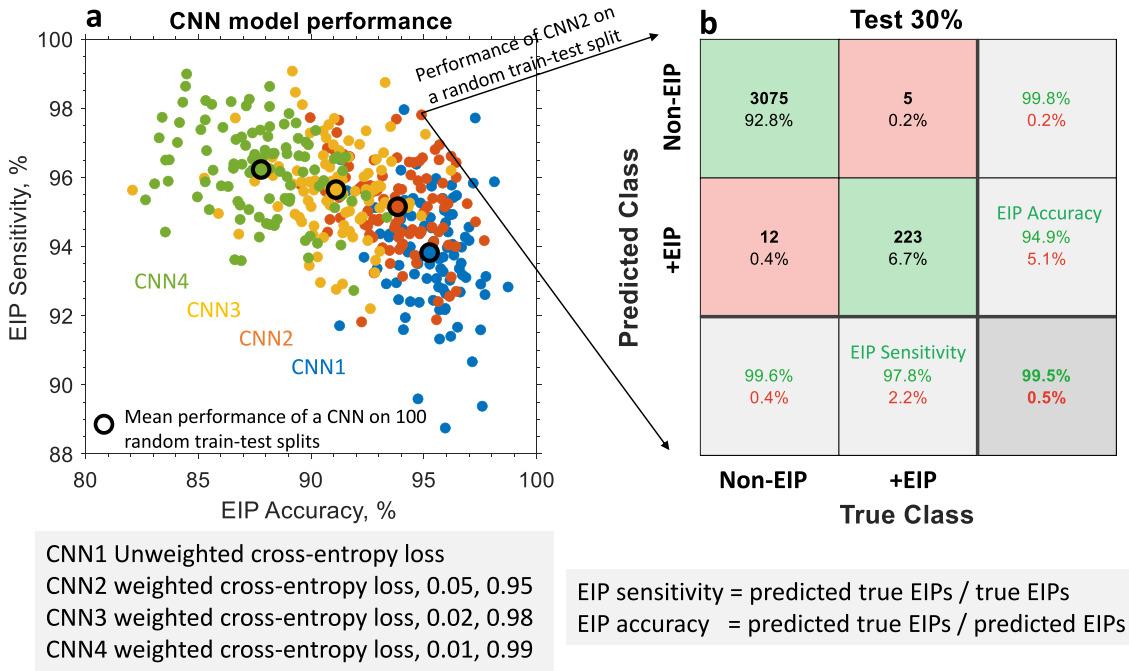

Figure 6. Classification performance of 4 CNN models on testing datasets using different weights of losses for EIP and non-EIP classes. (a) Performance of each CNN model running on 100 random train-test splits. (b) Confusion matrix (error matrix) for one single data point of CNN2 type.

weights on EIP and non-EIP classes in the loss function to overcome this issue. From CNN1 to CNN4, the model is more targeted to include as many true EIPs as possible with increased punishment on missing true EIPs, which also affects the accuracy. We use the standard definitions with the EIP sensitivity defined as the ratio of the model-predicted true EIPs over all true EIPs, and the EIP accuracy defined as the ratio of the model-predicted true EIPs over all model-predicted EIPs. The goal is to have both high EIP sensitivity and accuracy, but there is a tradeoff between the two measures.

Figure 6 shows the classification performance of the 4 CNN models. Each model was trained and tested on 100 random splits of the original >150 kA data set. There is a total of $4 \times 100 = 400$ data points in Figure 6a. These CNN classification models operate well with a mean EIP sensitivity of 95.2% and an EIP accuracy as high as 98.7%. Since the EIP sensitivity does not increase as significantly as the accuracy drops when the model puts more punishment on missing events, CNN2 appears to be an appropriate choice to use in practice. Figure 6b shows the confusion matrix (or error matrix) that visualizes the goodness of one example of training and testing result with CNN2. In this example, CNN2 was trained on $11{,}049 \times 70\% = 7{,}734$ events that were randomly selected from the entire data set, and then tested on the rest $11{,}049 \times 30\% = 3{,}315$ events. According to the confusion matrix, there are $223 + 5 = 228$ true EIPs in which 223 are identified by CNN2 but 5 are missed, while CNN2 predicted $223 + 12 = 235$ EIPs in which 223 are true but 12 are false. This means that CNN2 identified $223/228 = 97.8\%$ true EIPs with an accuracy of $223/235 = 94.9\%$, which is much better than unsupervised spectral clustering. Therefore, supervised CNN classification is demonstrated to be a powerful approach for +EIP classification and for lightning classification in general. In the next Section, we will apply the pretrained CNN model to search for +EIPs in a larger data set of >50 kA lightning events, aiming to detect more TGFs that were not feasible without ML.

3.3. Application in Searching for +EIPs > 50 kA

One of the pretrained CNN2 models is now applied to predict +EIPs from a new data set of 32,775 events of peak currents 50−150 kA. It should be noted that the relative fraction of component classes +EIPs, +NBEs, and +CGs in the data may not remain the same when the peak current threshold is lowered. Also, the new data set could include energetic pulses with a waveform pattern unlike any of the three categories identified in the high peak current data set, which could reduce classification performance. But we expect to be able to spot events that have a typical +EIP waveform shape using the pretrained CNN model.

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

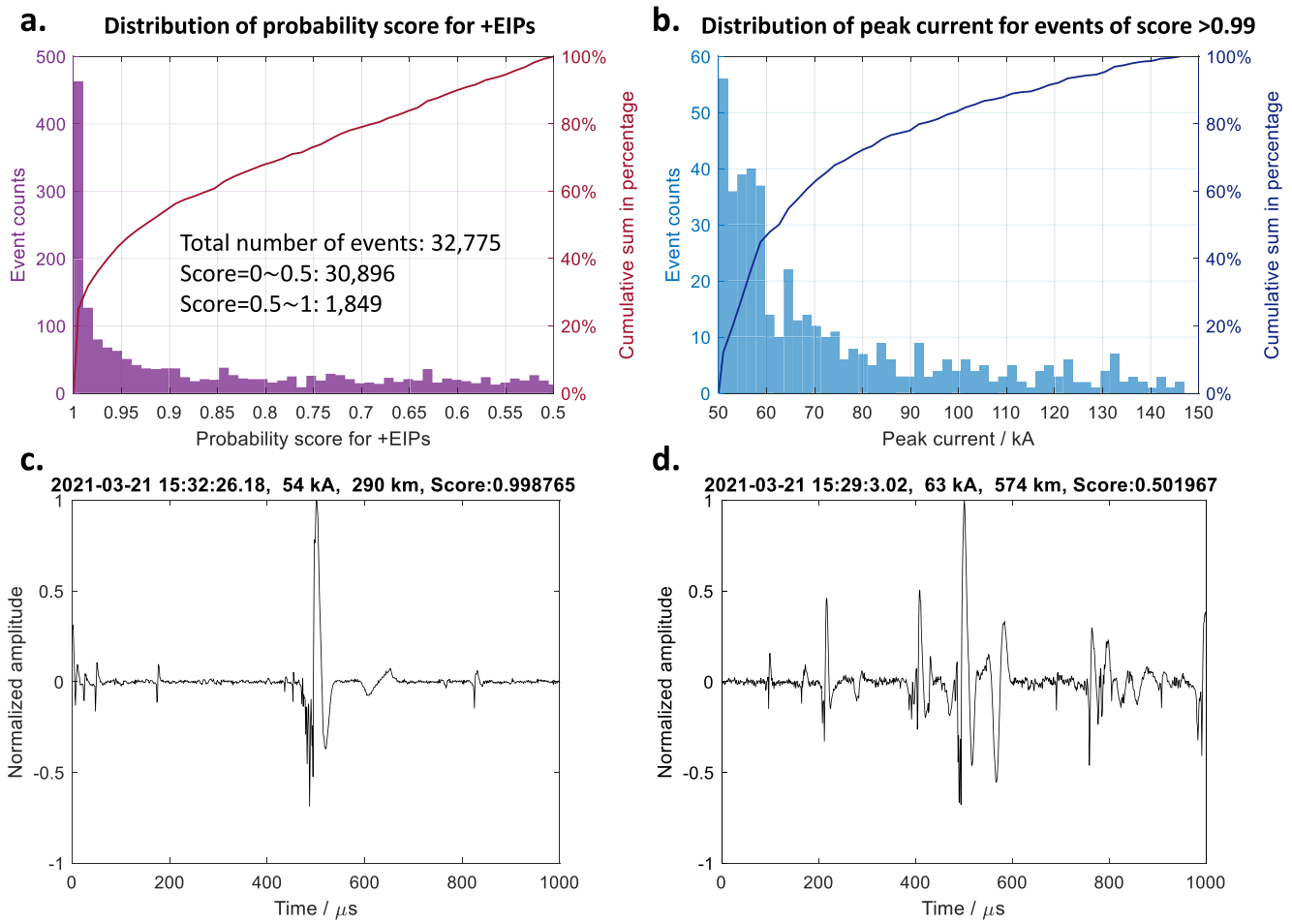

Figure 7. Prediction results of the pretrained CNN model on a new energetic lightning data set (~30,000 events) of peak currents between 50 and 150 kA. (a) Distribution of probability score for +EIPs. (b) Distribution of peak currents for events of probability score >0.99. (c) An example of lower peak current +EIPs identified by the CNN model with a high confidence close to 1. (d) Another example but with a low confidence close to 0.5.

Figure 7 shows the prediction results of the CNN model on the new 50−150 kA data set. 1,849 (5.6%) out of 32,775 events have a probability score higher than 0.5, which indicates that they are likely EIP events. The higher the probability score, the more likely the event is a true EIP. We then show two example waveforms of score = 0.999 and score = 0.502 in Figures 7c and 7d, respectively. The high-score event has extremely similar waveform to a typical >150 kA EIP, though its peak current is only 54 kA, while the low-score event contains more complicated pulses that make it difficult for unambiguous classification. Note that the CNN model has especially high confidence (score >0.99) in classifying 463 (23%) events into +EIPs, seeing the first high bar in the histogram plot of the probability score in Figure 7a. We thus assume that these 463 events are "true" +EIPs and proceed to investigate the peak current distribution of these lower peak current +EIPs (LEIPs), which is shown in Figure 7b. It is found that less than 20% LEIPs have a peak current above 100 kA while nearly half of them have relatively low peak current between 50 and 60 kA. This indicates that there could be many more EIP-TGFs happening in the thunderclouds than have been previously identified, if the EIP-TGF connection still holds for these LEIPs.

Then we investigate the Fermi-GBM photon data at the time of the lower peak current EIPs to verify if these LEIPs generated satellite-detectable TGFs. We slightly lower the score threshold to 0.9 to include more events to be our "true" +EIPs, which leads to 998 events. Since the Fermi satellite detects TGFs effectively within a distance of 600 km to its footprint (Briggs et al., 2013), only 10 LEIPs events are matched with Fermi within 600 km and 2 milliseconds. A case-by-case check on the LF waveforms and time-aligned photon counts of these 10 events shows that 2 are definitive TGFs that were previously reported by Fermi and 2 are suspected TGFs not

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

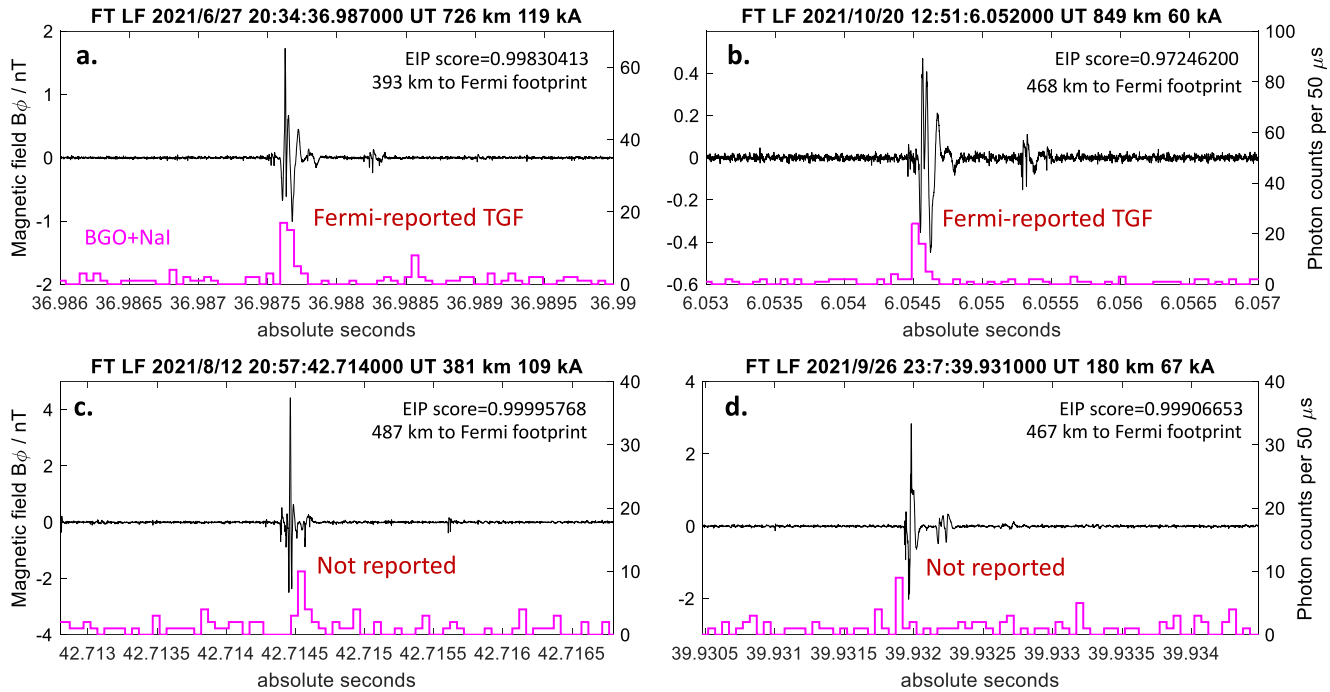

Figure 8. LF waveforms and Fermi gamma-ray counts for 4 lower peak current +EIPs identified by the CNN model. Two are Fermi-reported TGFs and two are suspect TGFs not reported by Fermi.

reported by Fermi but with a small peak (verified to not be a cosmic ray) higher than the environmental noise in the photon profile. These 4 cases are shown in Figure 8. However, the remaining 6 LEIPs are apparently not associated with detectable TGFs (see Figure S3 in Supporting Information S1), especially in one case that the Fermi satellite is almost right above the head with a small horizontal distance of 43 km but no excess of photon counts was detected at the time of this 52 kA LEIP. It is unknown whether this is because the TGF source is 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 collectively indicate that the relationship between TGFs and +EIPs of lower peak current (especially 50–100 kA) is complicated and needs to be further studied.

To this point, several questions arise that (a) What are those low-score energetic pulses (like the one in Figure 7d) and are they associated with TGFs? (b) What is the difference between a typical LEIP and a low-score energetic pulse? These questions call for further studies and we do not have definitive answers at the moment. However, the low-score pulses in Figure 7d also occur during the initial leader propagation similar to +EIPs and they are actually the so-called initial breakdown pulse (IBP) (e.g., Marshall et al., 2013; Y Wang et al., 2016; Wu et al., 2015). Some previous studies have suggested that strong IBPs could be candidates for TGFs (Marshall et al., 2013; Stolzenburg et al., 2016), and downward gamma rays (or weak TGFs) have been detected on the ground during IBPs of negative CGs (J Belz et al., 2018; J. W Belz et al., 2020; Wada et al., 2019).

3.4. Implications for Ionosphere Reflection From Clustering Results

As mentioned in Section 3.1, questions remain as to whether we can extract more information from Groups 1 to 3 in Figure 5c that seem to be separated by sky wave variation, and what the events inside Groups 4 and 5 are. Considering that the relationship between ground wave and sky wave is primarily reflected by the time difference of the two, it is thus important to align all events at the ground wave pulse instead of the absolute maximum of the entire pulse sequence. We then re-aligned lightning waveforms at the minimum of the first negative ground wave pulse and repeated the autoencoder feature extraction and spectral clustering.

In Figure 9, we show the group average waveforms for the 7 new groups, along with the distribution of event hour of the day and event distance to sensor for each group. Groups 6 and 7 again are primarily composed of +EIPs and +NBEs, respectively, making them relatively easy to discriminate from the rest groups. While ideally the number of events should be proportional to the square of distance (area), it is obvious from the distance distribution that

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

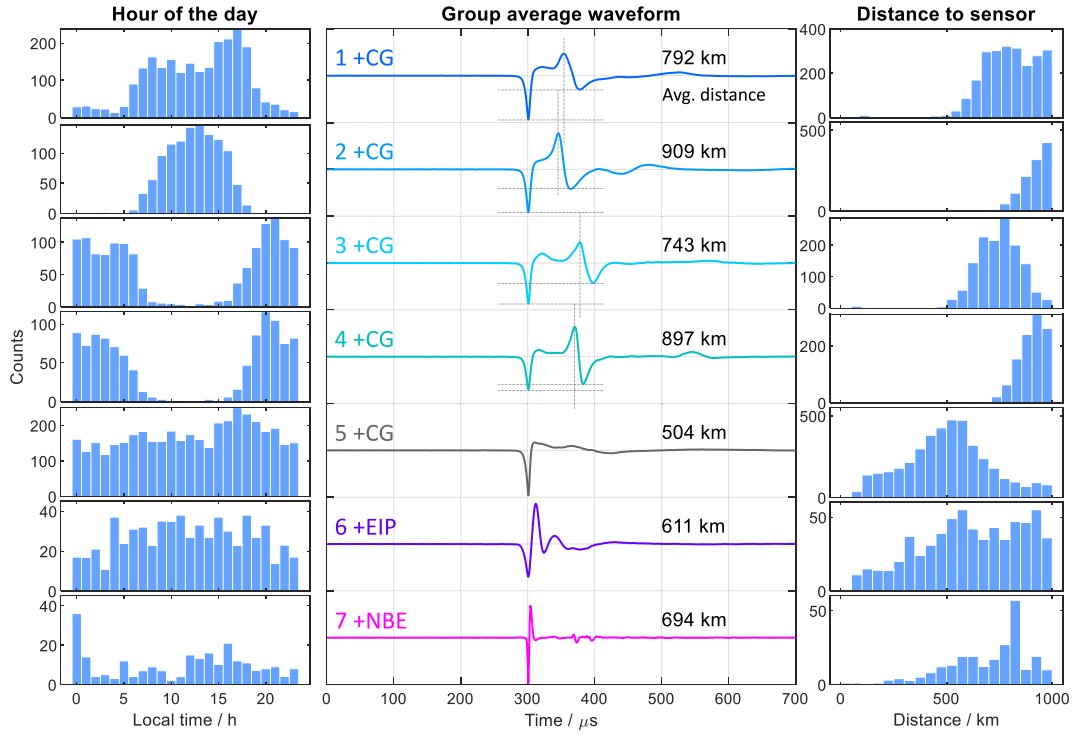

Figure 9. Group average waveforms and the corresponding histograms of the hour of the day and the distance to sensor for events in each group. The waveforms are aligned at the peak of the first ground wave pulse instead of the absolute maximum of the entire pulse as did in Figure 5.

the number of identified +EIPs stop increasing at a distance above 600 km, which again indicates that +EIPs become less distinguishable when they are far away from the sensor.

Compared with Figure 5c, the new Groups 1 to 5 become self-explanatory that they are all +CGs but with various sky waves. Note that we only see the first-hop sky wave in the group average waveforms, though there could have multiple reflected sky wave pulses in individual events due to multiple reflections. It is now clear that Groups 1 and 2 are mostly daytime +CGs, while Groups 3 and 4 nighttime +CGs. For daytime +CGs, comparing Group 2 with Group 1, the time difference between the ground wave and the sky wave is shorter, but the magnitude ratio of the sky-wave/ground-wave is larger. And the same case for nighttime events. Since Group 2(4) is farther from the sensor than Group 1(3), we can thus conclude that the sky-wave-to-group-wave time delay and the magnitude ratio are both related to the propagation distance. The reason is that the sky wave is generated by the reflection in ionosphere D region (e.g., Han & Cummer, 2010), which is sensitive to propagation distance. When the distance gets longer, the propagation length difference between the ground wave and the sky wave gets shorter and thus results in a shorter time difference between the two. Moreover, the ground wave attenuates faster than the sky wave, which leads to a relatively larger sky wave amplitude at a farther distance.

In addition, comparing nighttime events with daytime events, especially Group 2 and Group 4 that both occur near 900 km, we can see that the sky-wave-to-ground-wave time delay is longer for nighttime events, which indicates that the ionosphere reflection height is higher at night as can be explained by the reduced ionization by solar radiation. However, the sky-wave/ground-wave magnitude ratio is larger for nighttime events, which indicates that the ionosphere is more reflective during the night. Group 5 is relatively special that it does not contain obvious sky wave and the distribution of hour of the day is relatively uniform. Considering that most of events in Group 5 are within 600 km, it is plausible that close events do not have significant sky wave compared to ground wave, and this situation occurs during both the day and night.

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We can then summarize the characteristics of the ionosphere and its influence on the radio wave propagation based upon the ML unsupervised clustering results of energetic lightning radio waveforms:

1. The nighttime ionosphere reflection height is higher than the daytime.
2. The nighttime ionosphere reflection is stronger than the daytime.
3. The longer the propagation distance, the shorter the time delay between the sky wave and the ground wave.
4. The longer the propagation distance, the larger the magnitude ratio of sky-wave/ground-wave.
5. The sky wave of +CGs becomes significant at a distance above 500 km.

In fact, all these understandings align with the well-known knowledge about the ionosphere (e.g., Cummer et al., 1998; Shao & Jacobson, 2009; Wait, 1960), but the key point is that these characteristics are, to the best of our knowledge, shown for the first time to be revealed automatically by machine learning from lightning radio data. For the future, more detailed investigation on ionosphere remote sensing is anticipated with the help of machine learning.

3.5. Remarks on Broader Applications

To this point, we have demonstrated that the ML-based clustering and classification methods are extremely successful for revealing the underlying pattern of lightning radio signals and further classifying their types. In fact, the framework we proposed in Figure 3 can be used in any other waveform classification problems that first require initial understanding and labeling of the large data set, and then focus on identifying certain classes. There could be a lot of application scenarios beyond +EIPs and lightning classification, such as applications to earthquake signals, brain signals, radar signals, etc. Improvements in easing the process of the quality control of the initial labels from clustering results can be made by iterative training, prediction and refinement using CNNs.

4. Conclusions

Direct observation of TGFs with energetic particle detectors is powerful but also limited due to coverage in time and space, which in turn limits our understanding of TGF generation and effects. This work aims to improve the continuous and large-scale detection of TGFs through ground radio measurements of EIPs, based on the known EIP-TGF connection. As a manual search of extremely large volumes of lightning waveform data is infeasible, machine learning classifiers become a perfect tool to explore for this kind of big data classification problem.

First, we are interested in how many types of high peak current lightning pulses exist in the data and to confirm that +EIPs is a naturally distinct type, as has been shown previously (Lyu et al., 2015). For this purpose, we specially designed an unsupervised clustering model to automatically group different classes of energetic lightning events (>150 kA) independent of human knowledge. We implemented an autoencoder model to extract low-dimensional features from raw lightning waveforms and then applied spectral clustering to group different clusters for ∼10,000 events. As a result, the ML clustering shows that high peak current, positive NLDN polarity lightning pulses are composed of three distinct classes: +CGs (∼91%), +NBEs (∼2%) and +EIPs (6%–7%). This confirms that these three types are intrinsically different, and importantly, there is no sign of a fourth type of positive polarity high peak current lightning event.

High-performance EIP classification was then realized by a supervised CNN with a labeled data set generated by the unsupervised clustering. Since +EIPs are only 6%–7% of the data set, we specially tuned the weights of EIPs in the loss function for this scenario of imbalanced classes. Our CNN models identify on average 95.2% of true +EIPs with an accuracy up to 98.7%, making them a very powerful tool for automatic +EIP classification and possible ground-based TGF detection. In fact, such high sensitivity as well as the high accuracy is comparable to what a human can ideally achieve, and perhaps cannot be further improved because there is an intrinsic ambiguity of several percent in determining EIPs due to the complexity in individual waveforms.

We are also interested in how lower peak current EIP-like events are related to TGFs, knowing that previous definition of EIPs with a high threshold of 150 kA was arbitrary and largely due to the limited data volume processing capability of a human being. Thus, the pre-trained CNN classifier is then applied to a larger data set of ∼30,000 events with peak currents >50 kA. It is found that about a quarter of the predicted EIPs have relatively low peak currents between 50 and 60 kA, which suggests that many more TGFs could be happening than we already detected if the EIP-TGF connection still holds for LEIPs. Compared with Fermi-GBM photon data, 2 previously

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reported TGFs and 2 unreported but suspected TGFs are found out of 10 matched EIP-Fermi events, while the majority (6 cases) are apparently not associated with detectable TGFs. This indicates that LEIPs, especially those of peak currents between 50–100 kA, have a relatively complicated relationship with TGFs, but that at least some LEIPs are associated with TGFs detectable on orbit. Further observations and simulations are needed to investigate if there is a cutoff in EIP peak currents that controls/reveals the TGF generation and detectability at different source altitudes.

Meanwhile, similarities between LEIPs and IBPs are seen by ML because the CNN model predicts many IBPs to be EIPs with a probability score above 0.5. As the probability score increases, the ambiguity between EIPs and IBPs decreases. Note that there are a few reports of downward gamma rays during IBPs, it is worth a further study on the similarity and difference between the two types.

We additionally found that ML can automatically identify the diurnal change of the ionosphere reflection height and its effect on radio wave propagation. Note that waveform clustering results were affected by different time alignments attempted: aligning at absolute peak value and ground wave pulse peak. Absolute peak method was effective throughout the study to identify waveform types, but may not distinguish well between ground and sky waves. Ground pulse method was more complex, required human-defined criteria, but was beneficial for ionospheric feature extraction. In the future, we can extend the ML methods herein to investigate negative-polarity energetic lightning pulses and any other lightning sferics to extract more understandings hidden in the big lightning data. Moreover, the framework we proposed and demonstrated in this work for 1-D lightning waveform clustering and classification could be widely applied to other waveform classification problems, like earthquake signals, brain signals, radar signals, etc.

In summary, we develop machine learning classifiers, combining unsupervised and supervised methods, to automatically discriminate +EIPs with high sensitivity and accuracy, which represents a powerful tool for ground monitoring of EIP-type TGFs. Further application to lower peak current EIPs shows that the majority of 50–100 kA EIPs do not apparently relate to detectable TGFs.

Data Availability Statement

This work complies with the AGU data policy. The data set used in this study is available on the data repository website at https://doi.org/10.5281/zenodo.7651645 (Pu & Cummer, 2022). The ML model and figures were developed with MATLAB R2021a.

Acknowledgments
This study is supported by the National Science Foundation Dynamic and Physical Meteorology program through Grant AGS-2026304. The authors thank Amitabh Nag, Anjing Huang, Hamid Rassoul, Hamza Khounate, Mathieu Plaisir and Naomi Watanabe for their assistance with the LF system maintenance, and Melissa Gibby for help with the Fermi Data access. We also thank Simiao Ren and Joe Lucas for helpful discussions about machine learning.

References

Abbas, O. A. (2008). Comparisons between data clustering algorithms. *The International Arab Journal of Information Technology*, *5*(3).

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, *8*(1), 1–74. https://doi.org/10.1186/s40537-021-00444-8

Babich, L. P., Bochkov, E. I., Kutsyk, I. M., Neubert, T., & Chanrion, O. (2015). A model for electric field enhancement in lightning leader tips to levels allowing X-ray and gamma ray emissions. *Journal of Geophysical Research-Space Physics*, *120*(6), 5087–5100. https://doi.org/10.1002/2014ja020923

Belz, J., Abbasi, R., Le Von, R., Krehbiel, P., Remington, J., & Rison, W. (2018). High-energy particle showers observed at ground level in coincidence with downward lightning leaders at the Telescope Array Observatory. In *Paper presented at proceedings of 2016 international conference on ultra-high energy cosmic rays (UHECR2016)*.

Belz, J. W., Krehbiel, P. R., Remington, J., Stanley, M. A., Abbasi, R. U., LeVon, R., et al. (2020). Observations of the origin of downward terrestrial gamma-ray flashes. *Journal of Geophysical Research-Atmospheres*, *125*(23). https://doi.org/10.1029/2019jd031940

Briggs, M. S., Xiong, S., Connaughton, V., Tierney, D., Fitzpatrick, G., Foley, S., et al. (2013). Terrestrial gamma-ray flashes in the Fermi era: Improved observations and analysis methods. *Journal of Geophysical Research-Space Physics*, *118*(6), 3805–3830. https://doi.org/10.1002/jgra.50205

Celestin, S., & Pasko, V. P. (2011). Energy and fluxes of thermal runaway electrons produced by exponential growth of streamers during the stepping of lightning leaders and in transient luminous events. *Journal of Geophysical Research*, *116*(A3), 14. https://doi.org/10.1029/2010ja016260

Cummer, S. A., Inan, U. S., & Bell, T. F. (1998). Ionospheric D region remote sensing using VLF radio atmospherics. *Radio Science*, *33*(6), 1781–1792. https://doi.org/10.1029/98rs02381

Dwyer, J. R. (2003). A fundamental limit on electric fields in air. *Geophysical Research Letters*, *30*(20), 4. https://doi.org/10.1029/2003gl017781

Dwyer, J. R. (2012). The relativistic feedback discharge model of terrestrial gamma ray flashes. *Journal of Geophysical Research*, *117*(A2), 25. https://doi.org/10.1029/2011ja017160

Dwyer, J. R., Schaal, M. M., Cramer, E., Arabshahi, S., Liu, N., Rassoul, H. K., et al. (2012). Observation of a gamma-ray flash at ground level in association with a cloud-to-ground lightning return stroke. *Journal of Geophysical Research*, *117*(A10), 13. https://doi.org/10.1029/2012ja017810

Eack, K. B. (2004). Electrical characteristics of narrow bipolar events. *Geophysical Research Letters*, *31*(20), L20102. https://doi.org/10.1029/2004gl021117

Fishman, G. J., Bhat, P., Mallozzi, R., Horack, J., Koshut, T., Kouveliotou, C., et al. (1994). Discovery of intense gamma-ray flashes of atmospheric origin. *Science*, *264*(5163), 1313–1316. https://doi.org/10.1126/science.264.5163.1313

Grefenstette, B. W., Smith, D. M., Hazelton, B., & Lopez, L. (2009). First RHESSI terrestrial gamma ray flash catalog. *Journal of Geophysical Research*, *114*(A2), A02314. https://doi.org/10.1029/2008ja013721

Gulati, H., & Singh, P. (2015). Clustering techniques in data mining: A comparison. In *Paper presented at 2015 2nd international conference on computing for sustainable global development (INDIACom), IEEE*.

Han, F., & Cummer, S. A. (2010). Midlatitude daytime D region ionosphere variations measured from radio atmospherics. *Journal of Geophysical Research*, *115*(A10), 13. https://doi.org/10.1029/2010ja015715

Kohlmann, H., Schulz, W., & Pedeboy, S. (2017). Evaluation of EUCLID IC/CG classification performance based on ground-truth data. In *Paper presented at 2017 international symposium on lightning protection (XIV SIPDA), IEEE*.

Lu, G. P., Blakeslee, R. J., Li, J. B., Smith, D. M., Shao, X. M., McCaul, E. W., et al. (2010). Lightning mapping observation of a terrestrial gamma-ray flash. *Geophysical Research Letters*, *37*(11), 5. https://doi.org/10.1029/2010gl043494

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

Lyu, F. C., Cummer, S. A., Briggs, M., Marisaldi, M., Blakeslee, R. J., Bruning, E., et al. (2016). Ground detection of terrestrial gamma ray flashes from distant radio signals. *Geophysical Research Letters*, *43*(16), 8728–8734. https://doi.org/10.1002/2016gl070154

Lyu, F. C., Cummer, S. A., Briggs, M., Smith, D. M., Mailyan, B., & Lesage, S. (2021). Terrestrial gamma-ray flashes can be detected with radio measurements of energetic in-cloud pulses during thunderstorms. *Geophysical Research Letters*, *48*(11), e2021GL093627. https://doi.org/10.1029/2021gl093627

Marisaldi, M., Argan, A., Trois, A., Giuliani, A., Tavani, M., Labanti, C., et al. (2010). Gamma-ray localization of terrestrial gamma-ray flashes. *Physical Review Letters*, *105*(12), 128501. https://doi.org/10.1103/physrevlett.105.128501

Marshall, T., Stolzenburg, M., Karunarathne, S., Cummer, S., Lu, G. P., Betz, H. D., et al. (2013). Initial breakdown pulses in intracloud lightning flashes and their relation to terrestrial gamma ray flashes. *Journal of Geophysical Research-Atmospheres*, *118*(19), 10907–10925. https://doi.org/10.1002/jgrd.50866

Neubert, T., Østgaard, N., Reglero, V., Blanc, E., Chanrion, O., Oxborrow, C. A., et al. (2019). The ASIM mission on the international space station. *Space Science Reviews*, *215*(2), 1–17. https://doi.org/10.1007/s11214-019-0592-z

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

Ostgaard, N., Cummer, S. A., Mezentsev, A., Luque, A., Dwyer, J., Neubert, T., et al. (2021). Simultaneous observations of EIP, TGF, Elve, and optical lightning. *Journal of Geophysical Research-Atmospheres*, *126*(11), e2020JD033921. https://doi.org/10.1029/2020jd033921

Østgaard, N., Gjesteland, T., Stadsnes, J., Connell, P., & Carlson, B. (2008). Production altitude and time delays of the terrestrial gamma flashes: Revisiting the Burst and Transient Source Experiment spectra. *Journal of Geophysical Research*, *113*(A2), A02307. https://doi.org/10.1029/2007ja012618

Pu, Y. J., & Cummer, S. A. (2022). Unsupervised clustering and supervised machine learning for lightning classification: Application to identifying EIPs for ground-based TGF detection (version 2) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7651645

Pu, Y. J., Cummer, S. A., Lyu, F. C., Briggs, M., Mailyan, B., Stanbro, M., & Roberts, O. (2019). Low frequency radio pulses produced by terrestrial gamma-ray flashes. *Geophysical Research Letters*, *46*(12), 6990–6997. https://doi.org/10.1029/2019gl082743

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database Systems*, *5*, 532–538.

Roberts, O. J., Fitzpatrick, G., Stanbro, M., McBreen, S., Briggs, M. S., Holzworth, R. H., et al. (2018). The first fermi-GBM terrestrial gamma ray flash catalog. *Journal of Geophysical Research-Space Physics*, *123*(5), 4381–4401. https://doi.org/10.1029/2017ja024837

Shao, X. M., & Jacobson, A. R. (2009). Model simulation of very low-frequency and low-frequency lightning signal propagation over intermediate ranges. *IEEE Transactions on Electromagnetic Compatibility*, *51*(3), 519–525. https://doi.org/10.1109/temc.2009.2022171

Singh, A. (2010). Spectral clustering. *Machine Learning*, *10*(701/15), 781.

Stanley, M. A., Shao, X. M., Smith, D. M., Lopez, L. I., Pongratz, M. B., Harlin, J. D., et al. (2006). A link between terrestrial gamma-ray flashes and intracloud lightning discharges. *Geophysical Research Letters*, *33*(6), L06803. https://doi.org/10.1029/2005gl025537

Stolzenburg, M., Marshall, T. C., Karunarathne, S., & Orville, R. E. (2016). Luminosity with intracloud-type lightning initial breakdown pulses and terrestrial gamma-ray flash candidates. *Journal of Geophysical Research-Atmospheres*, *121*(18), 10919–10936. https://doi.org/10.1002/2016jd025202

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data*, *7*(1), 1–47. https://doi.org/10.1186/s40537-020-00349-y

Tilles, J. N., Krehbiel, P. R., Stanley, M. A., Rison, W., Liu, N., Lyu, F., et al. (2020). Radio interferometer observations of an energetic in-cloud pulse reveal large currents generated by relativistic discharges. *Journal of Geophysical Research: Atmospheres*, *125*(20), e2020JD032603. https://doi.org/10.1029/2020jd032603

Tran, M. D., Rakov, V. A., Mallick, S., Dwyer, J. R., Nag, A., & Heckman, S. (2015). A terrestrial gamma-ray flash recorded at the Lightning Observatory in Gainesville, Florida. *Journal of Atmospheric and Solar-Terrestrial Physics*, *136*, 86–93. https://doi.org/10.1016/j.jastp.2015.10.010

Tschannen, M., Bachem, O., & Lucic, M. (2018). Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069.

Ursi, A., Guidorzi, C., Marisaldi, M., Sarria, D., & Frontera, F. (2017). Terrestrial gamma-ray flashes in the BeppoSAX data archive. *Journal of Atmospheric and Solar-Terrestrial Physics*, *156*, 50–56. https://doi.org/10.1016/j.jastp.2017.02.014

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Wada, Y., Enoto, T., Nakazawa, K., Furuta, Y., Yuasa, T., Nakamura, Y., et al. (2019). Downward terrestrial gamma-ray flash observed in a winter thunderstorm. *Physical Review Letters*, *123*(6), 061103. https://doi.org/10.1103/PhysRevLett.123.061103

Wait, J. R. (1960). Terrestrial propagation of very-low-frequency radio waves, a theoretical investigation. *Journal of Research of the National Bureau of Standards*, *64*(2), 153. https://doi.org/10.6028/jres.064d.022

Wang, J., Huang, Q., Ma, Q., Chang, S., He, J., Wang, H., et al. (2020). Classification of VLF/LF lightning signals using sensors and deep learning methods. *Sensors*, *20*(4), 1030. https://doi.org/10.3390/s20041030

Wang, Y., Qie, X. S., Wang, D. F., Liu, M. Y., Su, D. B., Wang, Z. C., et al. (2016). Beijing Lightning Network (BLNET) and the observation on preliminary breakdown processes. *Atmospheric Research*, *171*, 121–132. https://doi.org/10.1016/j.atmosres.2015.12.012

Wu, T., Yoshida, S., Akiyama, Y., Stock, M., Ushio, T., & Kawasaki, Z. (2015). Preliminary breakdown of intracloud lightning: Initiation altitude, propagation speed, pulse train characteristics, and step length estimation. *Journal of Geophysical Research-Atmospheres*, *120*(18), 9071–9086. https://doi.org/10.1002/2015jd023546

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Xie, J., Girshick, R., & Farhadi, A. (2016). *Unsupervised deep embedding for clustering analysis, paper presented at International conference on machine learning*. PMLR.

Zhang, H. B., Lu, G., Lyu, F., Xiong, S., Ahmad, M. R., Yi, Q., et al. (2021). On the terrestrial gamma-ray flashes preceding narrow bipolar events. *Geophysical Research Letters*, *48*(8), e2020GL092160. https://doi.org/10.1029/2020gl092160

Zhu, Y. A., Bitzer, P., Rakov, V., & Ding, Z. Q. (2021). A machine-learning approach to classify cloud-to-ground and intracloud lightning. *Geophysical Research Letters*, *48*(1), e2020GL091148. https://doi.org/10.1029/2020gl091148

21698996, 2023, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022JD038369 by Duke University Libraries, Wiley Online Library on [03/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License