A Fast Radio Burst Dataset for Raw Data Searches

Authors: Xu Zhijun, An Tao, Guo ShaoGuang, Lao Baoqiang, Lü Weijia, Xiaocong Wu, Xu Zhijun

Date: 2022-06-28T00:00:00+00:00

Abstract

Fast radio bursts (FRBs) constitute an emerging frontier hotspot in international astronomy. The challenges posed by massive observational data in processing and analysis necessitate urgent research into intelligent search and identification of FRB signals. To accelerate FRB search research, we have developed a machine learning-based FRB dataset capable of training machine learning algorithms to detect FRBs in raw data. The current dataset comprises 8,020 FRB simulation images, 4,010 non-FRB images, and 4,010 radio frequency interference (RFI) simulation images, constructed based on open FRB observational results and expandable as needed. This study aims to provide an open-source dataset for state-of-the-art artificial intelligence algorithms to test and compare FRB identification methods. The dataset offers files in image and NumPy formats for both convolutional neural networks and classical machine learning algorithms, enabling binary classification between FRBs and non-FRBs, or ternary classification among FRBs, RFI, and background noise. In this example, we utilize 31 pre-trained classic convolutional neural networks (CNNs). For FRB/non-FRB classification, the models achieve 90-92% accuracy in the first training epoch, attaining a maximum accuracy of 99.8% on real data testing.

Full Text

A Machine Learning Dataset for Fast Radio Burst Detection in Raw Data

XU ZhiJun[12]*, AN Tao[12], GUO ShaoGuang[12], LAO BaoQiang[12], LU WeiJia[12] & WU XiaoCong[12]

[1] SKA Regional Centre Joint Lab, Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[2] SKA Regional Centre Joint Lab, Peng Cheng Lab, Shenzhen 518066, China

*Corresponding author, E-mail: xuthus@shao.ac.cn

Received date: ; **Accepted date:**
Funding: Supported by National Natural Science Foundation of China (Grant No. 12041301), National Key R&D Program Large Scientific Facility Frontier Research Project (2018YFA0404603), National Natural Science Foundation of China (Grant No. 11873079), and Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2021258)

Abstract

Fast radio bursts (FRBs) represent an emerging frontier in international astronomy, yet the massive observational data pose significant challenges for processing and analysis, necessitating intelligent search and identification of FRB signals. To accelerate FRB search research, we have developed a machine learning dataset that trains algorithms to detect FRBs directly in raw observational data. The dataset currently comprises 8,020 simulated FRB images, 4,010 non-FRB images, and 4,010 radio frequency interference (RFI) simulation images, all constructed from publicly available FRB observations and expandable as needed. This study provides an open-source dataset for state-of-the-art AI algorithms to test and compare FRB recognition methods. The dataset offers both image and NumPy format files for convolutional neural networks and classical machine learning algorithms, enabling either FRB/non-FRB classification or FRB/RFI/background noise classification.

Keywords: Fast Radio Burst, Machine Learning, Dataset
PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

1 Introduction

Fast radio bursts are bright millisecond-duration radio pulses [1,2]. Since their discovery in 2007, numerous radio telescopes including ASKAP, CHIME, and FAST have achieved significant observational breakthroughs, establishing FRBs as a hot topic in modern astronomy. Traditional FRB searches employ dedispersion methods: reading "filterbank" [3] or "FITS" [4] files from telescopes, removing radio frequency interference (RFI) [5,6], then searching through thousands of dispersion measures (DMs) from 100 to 2600 pc cm^{-3} to identify candidates [7,8]. Almost all FRB search pipelines [9–12] use conventional dedispersion algorithms for blind surveys. Despite many optimization efforts [13–16], these methods suffer from several drawbacks: extensive DM trials consume enormous computational resources; too many candidates require manual verification; and careful RFI removal is essential to avoid false positives.

We have generated a machine learning dataset for FRB detection in raw observational files. Unlike candidate-based searches [17,18], direct detection in raw data saves substantial computational time on dedispersion and RFI mitigation while enabling detection of weak FRB signals [19]. Furthermore, improving ML method accuracy through training can significantly reduce final candidate numbers.

To implement ML-based FRB searches, we developed the STEP software system[1], as ML accuracy heavily depends on training set quality. Currently, no large FRB dataset exists. This paper describes creating a dataset from ASKAP public data to improve and optimize models for raw data FRB searches. The dataset contains over ten thousand FRB images, simulated from 39 FRB signals detected by STEP in publicly available ASKAP data [20]. Figure 1 [Figure 1: see original paper] shows four FRB signals detected across two observations (in different beams). This dataset will be publicly released for the FRB community.

---

2 Dataset Construction

The FRB dataset is built from publicly available known FRB samples from ASKAP [21]. Construction involves three steps: first, detecting all known FRBs in raw data using conventional dedispersion pipelines; second, extracting these dedispersed FRB signals to build a simulation library; finally, constructing the raw-data-oriented FRB dataset by randomly selecting background data, FRB signals, and applying the methods and parameters described below (see Figure 2 [Figure 2: see original paper]).

2.1 FRB Observations and Search

We utilized data from [21], which released 19 FRB observations, each containing 36-beam "filterbank" files. The data are 8-bit with 336 channels of 1 MHz each, sampled at 1.26 ms, sorted in lower sideband, with a maximum frequency of 1488 MHz[2].

We used STEP, our in-house GPU-based open-source toolkit for FRB search and analysis, developed and tested on the China SKA Regional Centre Prototype (CSKA-P) [22]. STEP successfully detected all known FRB signals from public ASKAP data, processing 36 beams of 3,295-second observations in just 2,743 seconds on a single GPU.

2.2 Simulating FRB Signals

As FRB signal characteristics continue to be discovered and analyzed, the best simulation approach uses real FRB signals as templates. Therefore, we use STEP to search for and extract dedispersed FRB signals, then simulate new signals using the methods and parameters below before injecting them into real observational background data to generate dataset samples.

2.2.1 Dispersion Measure Dispersion measure is a primary FRB characteristic that determines the delay between highest and lowest frequencies. Known FRB DMs range from 100 to 2600 pc cm^{-3} [7,8], but searching for more distant exotic FRBs benefits from larger ranges, so the maximum can exceed 2600 pc cm^{-3}. The minimum of 100 pc cm^{-3} is typically maintained as it corresponds to pulsar or RFI signal ranges. In our simulations, FRB DMs are randomly selected from 100 pc cm^{-3} to a maximum determined by bandwidth, frequency, sampling time, and image pixels (see Section 4.1).

Since DM has a linear relationship with delay, we can extend support for larger DMs by preprocessing raw data with specified DM values. For example, if images support a maximum DM of 400, DMs exceeding 400 require preprocessing with DM=300 dedispersion (the 100 DM offset exists because DMs below 100 are outside the search range), raising the supported maximum to 700. This process can be repeated for even higher DMs. This approach allows custom minimum and maximum DM ranges based on local computational resources. Note this preprocessing method only applies during inference; it is not needed for dataset generation or training, but explains how to extend DM ranges for real observations.

2.2.2 RFI Signals Simulated FRB signals are injected into raw data that retains background RFI. ASKAP's site conditions are excellent, making RFI rare in most data, so we manually inject simulated RFI signals. In our simulations, RFI signals are assigned zero or negative DM, constraining the ML algorithm to identify only positive DMs greater than 100 pc cm^{-3}, consistent with real observations.

2.2.3 Fluence Observations show FRB energy and fluence vary dramatically. Even for repeaters, different pulses exhibit different fluences, and the same pulse shows varying fluence across beams (see Figure 1). This indicates that adjusting simulated signal fluence within appropriate ranges not only expands sample diversity but also better approximates reality. To enhance detection of weak FRBs, we randomly amplify or attenuate signal fluence from near-detection limits to actual observed FRB fluence values.

2.3 Dataset Statistics

ASKAP FRB observations contain 36 beams. Beams with known FRB signals are identified, allowing beams without FRBs to serve as noise backgrounds for generating non-FRB images. Injecting RFI and FRB signals into these backgrounds creates RFI and FRB images. The complete dataset contains 16,040 images: 4,010 raw data (no FRB), 4,010 RFI, 4,010 FRB, and 4,010 weak FRB images (see Figure 3 [Figure 3: see original paper]). Real FRB signals for simulation come from four beams of FRB170906 (see Figure 4 [Figure 4: see original paper]), while background data come from other beams in the same observation

[21] without FRB signals, also used for injecting signals to simulate FRB or RFI.

Dataset formats include images and NumPy files. Image formats can be "ps", "pdf", "svg", or default "png" supported by CNN algorithms. STEP supports NumPy "npy" format or conversion to filterbank files compatible with other FRB search pipelines. It also supports classical ML algorithms (e.g., SVM, Random Forest), enabling performance comparisons between CNNs, classical ML, and traditional FRB pipelines.

---

3 Experiments

Experiments were conducted on the China SKA Regional Centre Prototype [23], which features a comprehensive software platform [24], high-speed data transmission from SKA precursor facilities [25,26], and optimized radio astronomy pipelines [27,28]. The hardware comprises 4 GPU nodes with 16 NVIDIA V100 GPUs and 4 A40 GPUs. This study used 3 GPU nodes with 16 NVIDIA V100 GPUs for training.

3.1 Model Architecture

We developed models to test the FRB dataset using PyTorch[3] [29] and its torchvision package, which includes modern datasets, popular model architectures, and standard image transformations. The PyTorch 1.6.0 model package contains the following image classification algorithms with pre-trained models: - VGG (vgg11, vgg13, vgg16, vgg19, vgg11_{bn}, vgg13_{bn}, vgg16_{bn}, vgg19_{bn}) [30] - DenseNet (densenet121, densenet169, densenet201, densenet161) [31] - ResNet (resnet18, resnet34, resnet50, resnet101, resnet152) [32] - ResNeXt (resnext50_{32x4d}, resnext101_{32x8d}) [33] - Wide ResNet (wide_{resnet50}2, wide_{resnet101}2) [34] - AlexNet [35], inception_{v3} [36], GoogLeNet [37], mobilenet_{v2} [38] - ShuffleNet v2 (shufflenet_{{{v2}}*{{x0}}}5, shufflenet*{{{v2}}_{{x1}}}0) [39] - SqueezeNet (squeezenet1_0, squeezenet1_1) [40] - MNASNet (mnasnet0_5, mnasnet1_0) [41]

3.2 Pre-trained Models

Conventional training initializes networks randomly, which is time-consuming and unstable. This study employs pre-trained models trained on massive datasets [42] for initialization. Since our dataset differs significantly from the original and may be large, we fine-tune the entire network after pre-training.

3.3 Training

We split the dataset 8:2 into 12,832 training images and 3,208 validation images. We evaluated 31 classic CNN algorithms with pre-trained models. Using pre-trained models, most networks achieved 90-92% accuracy in the first epoch, exceeding 99.7% after several epochs. This demonstrates that classic CNNs effectively extract features from simulated FRB images and improve accuracy through iterative training. To comprehensively evaluate classification performance, we computed recall, precision, and F1 scores. Recall reflects positive sample identification capability, precision indicates false positive probability, and F1 provides a balanced metric. Table 1 shows most models achieve recall, precision, and F1 scores above 99%, confirming our simulated FRB dataset is highly suitable for classic CNN algorithms.

3.4 Test Dataset

To validate CNN models trained on simulated data, we generated 16,544 real FRB signal images from all public ASKAP FRB observations as a test dataset, including 35 images with FRB signals and the remainder as non-FRB images. The data extracts approximately 10 seconds of raw data containing FRB signals, matching real observation processing or candidate buffer data from transient source terminals. Table 2 shows most networks achieve 99% accuracy on real data, with deeper networks generally performing better. In practice, maximizing recall is crucial to minimize missing FRBs (positive samples). Table 2 shows three models achieve 100% recall, detecting all FRB signals. This 100% recall occurs because positive FRB samples constitute a tiny fraction of real data, making recall potentially imprecise with limited positives and highlighting the importance of high recall in real scenarios. With 100% recall, the highest precision is only 99.321%, indicating some false positive candidates requiring subsequent manual review.

---

4 Discussion

4.1 Image Size and Signal Position in Real Data

A primary dataset limitation is image size. CNN algorithms typically require fixed-size square images. However, in real FRB data, image height is defined by channel count and width by sample duration. Channel counts are usually fixed, so we must select time samples equal to channel count for square images.

Our generation algorithm assumes all frequency components of an FRB appear within one image. For real data, this requires constraints on signal position and DM. For unknown FRBs, we use overlapping regions to satisfy position constraints. For ASKAP data, we set 50% overlap: if a single image spans time length 1, new images are generated every 0.5 time units. This ensures all frequency components fall within single images for certain DM ranges.

Regarding DM constraints, as noted in Section 2.2.1, DM is proportional to delay, so maximum DM is limited by sample duration. While temporal downsampling could solve this, it risks sensitivity loss. Instead, we pre-dedisperse high-DM signals. For ASKAP data, we set a pre-dedispersion step of 400 pc cm^{-3}. Searching up to 2000 pc cm^{-3} requires five parallel pipelines: no pre-dedispersion and pre-dedispersion at 400, 800, 1200, and 1600 pc cm^{-3}. Each pipeline then searches only within 400 pc cm^{-3}, ensuring all signals fit within single images and improving search accuracy. While removing the single-image constraint could fundamentally solve this limitation, it introduces other challenges requiring future investigation.

4.2 Weak FRB Signals

This issue arose from early errors in random fluence transformations. Initially, we applied randomly selected fluence factors to all frequency components simultaneously. Upon reviewing generated images, we observed "anomalous weak lines" (see Figure 3) that do not appear in real data. This occurs because FRB signals exhibit frequency-dependent fluence; random scaling can make some frequency components weaker than background, creating these artifacts. We resolved this by checking injected signals and replacing components weaker than background noise with background values, making simulations more realistic.

4.3 FRB Width

FRB signal width determines how many real FRB samples must be extracted from observations. Maximum and minimum FRB widths remain uncertain, requiring manual extraction of real FRB signals. While width could be randomized, our simulations currently use fixed widths.

---

5 Future Work and Conclusion

The dataset represents only the first step in FRB searches. Future work includes incorporating observations and RFI data from different telescopes to improve adaptability. For SKA Phase 1 FRB studies, we can use MWA data from the SKA-low site to develop SKA-low FRB simulation datasets, and MeerKAT data for SKA-mid simulations. Additionally, FRB energy variations across frequencies and quantization effects in different data formats require telescope-specific studies. A major challenge in FRB pipeline development is the lack of standardized comparison metrics. Our simulation dataset enables pipeline comparison and optimization, providing unified standards for development, testing, and performance quantification.

To accelerate FRB search pipeline research, we developed a machine learning dataset that trains algorithms to detect FRBs in raw data while serving as a performance benchmark for traditional pipelines. The dataset currently contains 8,020 FRB, 4,010 non-FRB, and 4,010 RFI images constructed from public observations. We provide this open-source dataset for state-of-the-art AI algorithms to compare FRB identification methods. The dataset offers image and NumPy formats for CNN and classical ML algorithms, supporting FRB/non-FRB or FRB/RFI/background classification. The image dataset is already open-source; the simulation toolkit will be released after further testing to meet specific telescope search requirements.

References

1. Lorimer D R, Bailes M, McLaughlin M A, et al. A Bright Millisecond Radio Burst of Extragalactic Origin. Science, 2007, 318(5851): 777–780
2. Petroff E, Hessels J W T, Lorimer D R. Fast radio bursts. Astron Astrophys Rev, 2019, 27(1): 1–75
3. Lorimer D R. SIGPROC: pulsar signal processing programs. Astrophysics Source Code Library, 2011, ascl: 1107.016
4. Hotan A W, Van Straten W, Manchester R N. PSRCHIVE and PSRFITS: an open approach to radio pulsar data storage and analysis. Publ Astron Soc Aust, 2004, 21(3): 302–309
5. Petroff E, Keane E F, Barr E D, et al. Identifying the source of Perytons at the parkes radio telescope. Mon Notices Royal Astron Soc, 2015, 451(4): 3933–3940
6. Nita G M, Gary D E. The generalized spectral kurtosis estimator. Mon Notices Royal Astron Soc, 2010, 406(1): L60–L64
7. Bhandari S, Keane E F, Barr E D, et al. The SUrvey for Pulsars and Extragalactic Radio Bursts–II. New FRB discoveries and their follow-up. Mon Notices Royal Astron Soc, 2018, 475(2): 1427–1448
8. Caleb M, Keane E F, Van Straten W, et al. The SUrvey for Pulsars and Extragalactic Radio Bursts–III. Polarization properties of FRBs 160102 and 151230. Mon Notices Royal Astron Soc, 2018, 478(2): 2046–2055
9. Ransom S. PRESTO: PulsaR Exploration and Search TOolkit. Astrophysics source code library, 2011, ascl: 1107.017
10. Champion D J, Petroff E, Kramer M, et al. Five new fast radio bursts from the HTRU high-latitude survey at Parkes: first evidence for two-component bursts. Mon Notices Royal Astron Soc, 2016, 460(1): L30–L34
11. Bannister K W, Shannon R M, Macquart J P, et al. The detection of an extremely bright fast radio burst in a phased array feed survey. Astrophys J Lett, 2017, 841(1): L12
12. Mikhailov K, Sclocco A. The Apertif Monitor for Bursts Encountered in Real-time (AMBER) auto-tuning optimization with genetic algorithms. Astronomy and computing, 2018, 25: 139–148
13. Barsdell B R, Bailes M, Barnes D G, et al. Accelerating incoherent dedispersion. Mon Notices Royal Astron Soc, 2012, 422(1): L57–L61
14. Sclocco A, van Leeuwen J, Bal H E, et al. Real-time dedispersion for fast radio transient surveys, using auto tuning on many-core accelerators. Astronomy and Computing, 2016, 14: 1–14
15. Zackay B, Ofek E O. An accurate and efficient algorithm for detection

of radio bursts with an unknown dispersion measure, for single-dish telescopes and interferometers. Astrophys J, 2017, 835(1): 11

16. Amiri M, Bandura K, Berger P, et al. The CHIME fast radio burst project: system overview. Astrophys J, 2018, 863(1): 48
17. Connor L, van Leeuwen J. Applying deep learning to fast radio burst classification. Astron J, 2018, 156(6): 256
18. Agarwal D, Aggarwal K, Burke-Spolaor S, et al. FETCH: A deep-learning based classifier for fast transient classification. Mon Notices Royal Astron Soc, 2020, 497(2): 1661–1674
19. Zhang Y G, Gajjar V, Foster G, et al. Fast radio burst 121102 pulse detection and periodicity: a machine learning approach. Astrophys J, 2018, 866(2): 149
20. McConnell D, Allison J R, Bannister K, et al. The Australian Square Kilometre Array Pathfinder: Performance of the Boolardy Engineering Test Array. Publ Astron Soc Aust, 2016, 33: e042
21. Shannon R M, Macquart J P, Bannister K W, et al. The dispersion–brightness relation for fast radio bursts from a wide-field survey. Nature, 2018, 562(7727): 386–390
22. An T, Wu X P, Hong X. SKA data take centre stage in China. Nat Astron, 2019, 3(11): 1030–1030
23. An T, Wu X C, Lao B Q, et al. Status and progress of China SKA Regional Centre prototype. arXiv:2206.13022
24. Lao B Q, Zhang Y K, An T, et al. Software Platform on China SKA Regional Center Prototype System (in Chinese). ChinaXiv:202206.00173
25. Guo S G, An T, Xu Z J, et al. Progress and Prospect of transcontinental high-speed data transmission at SKA Regional Center in China (in Chinese). ChinaXiv:xxxx
26. Guo S G, Lu Y, An T, et al. Scientific data flow and array simulation analysis for the SKA-1 era (in Chinese). ChinaXiv:xxxx
27. Wei J W, Zhang C F, Zhang Z L, et al. Parallel optimization of the pulsar search pipeline on China SKA Regional Centre Prototype (in Chinese). ChinaXiv:T202206.00297
28. Wei J W, Zhang C F, Lao B Q, et al. Optimization of parallel processing of Square Kilometre Array low frequency imaging pipeline (in Chinese). ChinaXiv:T202206.00292
29. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019, 32: 8026-8037
30. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115(3): 211–252
31. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700–4708
32. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778

33. Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492–1500
34. Zagoruyko S, Komodakis N. Wide residual networks. arXiv:1605.07146
35. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25: 1097–1105
36. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818–2826
37. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1–9
38. Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
39. Zhang X, Zhou X, Lin M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848–6856
40. Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv:1602.07360
41. Tan M, Chen B, Pang R, et al. MnasNet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2820–2828
42. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556

---

[1] https://github.com/Xu-Zhijun/STEP
[2] Shannon, Ryan; Bannister, Keith (2018): Data from the ASKAP latitude 50 Fast Radio Burst (FRB) sample. v3. CSIRO. Data Collection. https://doi.org/10.25919/5b6ae6b515850
[3] https://pytorch.org/

Note: Figure translations are in progress. See original paper for figures.