DATASHEET:

Jefferson Lab C100 Superconducting Radio-Frequency Cavity Fault Data, 2020

Adam Carpenter
Thomas Jefferson National Accelerator Facility
adamc@jlab.org

Tom Powers
Thomas Jefferson National Accelerator Facility
powers@jlab.org

Anna Shabalina
Thomas Jefferson National Accelerator Facility
shabalin@jlab.org

Chris Tennant
Thomas Jefferson National Accelerator Facility
tennant@jlab.org

Lasitha Vidyaratne
Thomas Jefferson National Accelerator Facility
lasithav@jlab.org

This document is based on *Datasheets for Datasets* by Gebru et al. [1]. Please see the most updated version here.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to train machine learning models for the task of identifying the (1) cavity and (2) fault type from C100 cryomodules at Thomas Jefferson National Accelerator Facility (Jefferson Lab), thereby replacing the time-consuming efforts of a subject matter expert. Superconducting radio-frequency (SRF) cavity trips represent a significant source of machine downtime. Real-time – rather than post-mortem – identification of the offending cavity and classification of the fault type would give control room operators valuable feedback for corrective action planning. The anticipated benefit is increased beam-on-target time for users and provides performance metrics that can be used to improve future cavity designs.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Adam Carpenter, Tom Powers, Chris Tennant and Lasitha Vidyaratne, representing the Accelerator Division at Thomas Jefferson National Accelerator Facility.

What support was needed to make this dataset? (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

This work was supported by Department of Energy (DOE) contract DE-AC05-06OR23177, under which Jefferson Science Associates, LLC, operates the Thomas Jefferson National Accelerator Facility.

Any other comments?

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each instance corresponds to features extracted from time-series data captured during a C100 SRF cavity fault.

How many instances are there in total (of each type, if appropriate)?
2,375

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
Data is recorded and written to file every time a C100 cavity trips, using a specially designed data acquisition system. The dataset represents fault events recorded during CEBAF beam operations between January 18, 2019 and March 9, 2020. By default, the data acquisition system records 17 RF signals per cavity. However, system experts identified the 4 signals that have the highest predictive power. The features extracted from these select signals constitute the dataset. During data preprocessing some events are neglected because they are missing data or are otherwise corrupt (see PREPROCESSING / CLEANING / LABELING section).

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance includes a timestamp ("date_time"), a label for the cryomodule which experienced the trip ("zone_label"), and 192 features ("feature_1", "feature_2"... "feature_192"). When a C100 SRF cavity fault occurs, RF signals from each of the 8 cavities that comprise a C100 cryomodule are written to file. The features correspond to 6 autoregressive features for each of 4 signals per cavity for each of the 8 cavities (6×4 signals/cavity $\times$ 8 cavities/cryomodule = 192).

Is there a label or target associated with each instance? If so, please provide a description.
For each instance, there is an associated label for the (1) cavity which faulted first ("cavity_labels.csv") and (2) the type of fault that caused the trip ("fault_labels.csv"). The cavity identification can take a values [0,1,2,3,4,5,6,7,8] and the fault type can take values of ['Microphonics', 'Quench_100ms', 'Controls_Fault', 'E_Quench', 'Quench_3ms', 'Single_Cav_Turn_Off' , 'Heat_Riser_Choke', 'Multi_Cav_Turn_Off']. For a technical discussion about fault types see Ref. [3].

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
A common set of features ('features.csv') is used to train machine learning models for both cavity identification ('cavity_labels.csv') and fault classification ('fault_labels.csv'). Each model was trained using a random, stratified split (70% training, 30% testing) of the appropriate labels.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
Redundancies and incomplete data were removed in preprocessing (see "PREPROCESSING / CLEANING / LABELING" section).

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Any other comments?

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The features in the dataset are computed from time-series signals collected by a specially designed data acquisition system. When a C100 cavity faults, data is automatically written to file.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The dataset represents C100 cavity faults collected from January 18, 2019 to March 9, 2020. The dataset was first published April 30, 2020.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Each of the eight cavities within a C100 cryomodule has an independent Field Control Chassis (FCC) EPICS Input/Output Controller (IOC), capable of buffering and presenting 17 diagnostic waveforms. These waveforms are stored in buffers containing 8192 points and sampled at a configurable rate typically ranging from 5 kHz to 20 kHz. Upon experiencing an RF fault, these buffers are frozen and made available in an EPICS waveform and neighboring IOCs are rapidly notified of the fault. This allows for a time-synchronized set of waveforms to be produced across all cavities within the cryomodule. In addition to the raw waveform data, the IOCs also present PVs for the timestamp associated with the fault, the sampling interval, and the relative offset of fault from the start of the buffered waveform. (See Ref. [4]).

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
After the data acquisition system was designed and implemented, routine software maintenance was performed by laboratory staff.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Any other comments?

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Cleaning

Several checks were implemented to filter the raw data.

- Only events in which all 17 RF signals for each of the 8 cavities are written to file were retained
- Only fault events with the data sampled at 5 kHz were retained
- Fault data originating from cryomodule 0L04 were neglected
- Any data collected between 3PM on February 4, 2019 and 12PM on February 5, 2019 were ignored
- A small number of instances were labeled multiple times by the same labeler with differing labels. These were ignored.
- Events were ignored if faults did not occur when all operating cavities were in generator driven resonance (GDR) mode for RF operations. GDR mode is a prerequisite for beam operation. Non-GDR mode faults typically occurred during recovery from a recent fault, but may happen for a variety of reasons. This is considered a reasonable data cut, so that the data contains only faults during beam operation.

Preprocessing

The values in the raw signal waveforms feature orders-of-magnitude variations among cavities and among signal types within single cavities. Therefore, prior to feature extraction time-series standardization is applied to each waveform (see Ref. [5]).

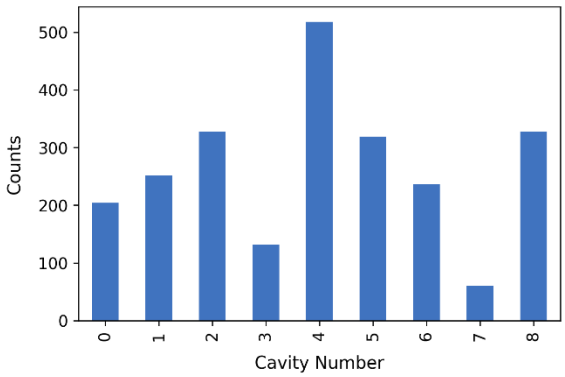
Labeling

Labeling of the data (assigning the cavity identification and fault classification corresponding to each event) was done by Tom Powers, a subject matter expert with over 30 years of experience with Jefferson Lab SRF systems.

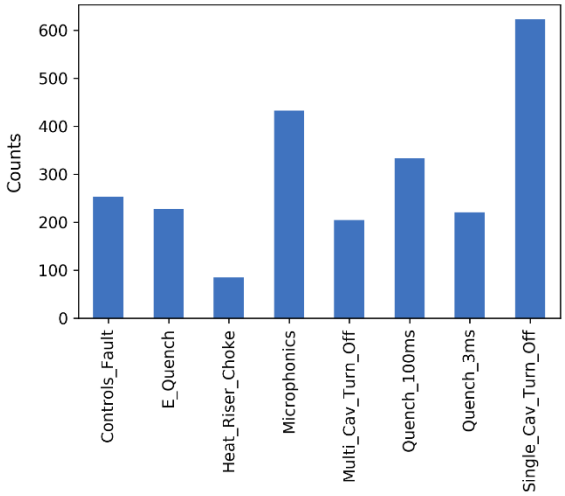
Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Yes, the raw data is compressed and archived at Thomas Jefferson National Accelerator Facility. It is not publicly accessible.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
No.

Any other comments?
A breakdown of the events by cavity number label is shown in the figure below:



A breakdown of the events by fault type label is shown in the figure below:



USES

Has the dataset been used for any tasks already? If so, please provide a description.
The dataset was used to train and evaluate machine learning models for C100 SRF cavity and fault identification at CEBAF.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
As of April 30, 2020 there have been no papers that utilize the dataset.

What (other) tasks could the dataset be used for?
Using clustering methods (unsupervised learning) one may be able to identify the unique number fault types from the features of the dataset, rather than relying on a subject matter expert.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks). If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
The dataset is specific to Thomas Jefferson National Accelerator Facility's C100-style SRF cavities and more broadly, to the operating conditions of the CEBAF.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Models trained on the dataset will not generalize well to other SRF accelerator facilities.

Any other comments?

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset is publicly accessible (see below). There are no plans to actively distribute the data except to colleagues joining our collaboration (i.e. graduate students).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is publicly accessible as a GitHub repository:
https://github.com/JeffersonLab/AI_SRF_operations/tree/master/datasets/C100-2020-04-30
DOI: 10.14462/MLFaultClassifier/1616675

When will the dataset be distributed?
The dataset was made available on the GitHub repository April 30, 2020.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

Any other comments?

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
Adam Carpenter and Chris Tennant

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
adamc@jlab.org, tennant@jlab.org

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
There are no plans to add or delete instances to the dataset. As more data is collected, additional, separate datasets will be created.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
We do not anticipate the current version of the dataset evolving (no instances will be added or deleted). It will be maintained on a publicly accessible repository (GitHub).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
No.

Any other comments?

REFERENCES

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wort-man Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. arXiv:1803.09010 [cs], January 2020.

[2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243 [cs], June 2019.

[3] T. Powers and A. Solopova Shabalina, “CEBAF C100 Fault Classification based on Time Domain RF Signals”, presented at the 19th Int. Conf. RF Superconductivity (SRF'19), Dresden, Germany, paper WETEB3 (2019).

[4] Adam Carpenter, et al., “Initial Implementation of a Machine Learning System for SRF Cavity Fault Classification at CEBAF.”17th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS’19), New York, NY, USA, Oct. 2019, paper WEPHA025.

[5] Lasitha Vidyaratne, “Machine Learning Pipeline for C100 Cavity and Fault Classification”, Jefferson Lab Technical Note 20-014 (2020).