A DATA-DRIVEN ANOMALY DETECTION ON SRF CAVITIES AT THE EUROPEAN XFEL

A. Sulc*, A. Eichler, T. Wilksen, DESY, Hamburg, Germany

Abstract

The European XFEL is currently operating with hundreds of superconducting radio frequency cavities. To be able to minimize the downtimes, prevention of failures on the SRF cavities is crucial. In this paper, we propose an anomaly detection approach based on a neural network model to predict occurrences of breakdowns on the SRF cavities based on a model trained on historical data. We used our existing anomaly detection infrastructure to get a subset of the stored data labeled as faulty. We experimented with different training losses to maximally profit from the available data and trained a recurrent neural network that can predict a failure from a series of pulses. The proposed model is using a tailored architecture with recurrent neural units and takes into account the sequential nature of the problem which can generalize and predict a variety of failures that we have been experiencing in operation.

INTRODUCTION

The superconducting radio-frequency (SRF) cavities are responsible for accelerating beams which are used in the European X-ray Free Electron Laser (EuXFEL) to obtain extremely brilliant X-ray photon light.

Particle accelerators use the cavity resonators operating in radio-frequency spectra to accelerate particles by synchronization with their frequency. The cavities accelerate and energize particles by the induced alternating electric field.

For superconductivity, it is necessary to maintain the cavities cooled to very low temperatures, usually near absolute zero, with a cryogenic system. The cryogenic system maintains the temperature to preserve the superconductivity. The superconductivity minimizes the losses through the wall to a minimum and thus almost all RF power can be transmitted to the passing beam.

EuXFEL is currently operating 784 SRF cavities and it is necessary to use automated algorithms to prevent failures. One kind of failure we are particularly interested in are quenches. A quench is when cavity walls lose their superconductivity due to temperature breakdown. It leads to a loss of superconductivity and energy is dissipated through the cavity walls (the surrounding helium bath is heated up) thus the quality factor decreases, i.e. efficiency decreases. Although the quench limits are experimentally tested and set in the firmware to hard limits, there are numerous situations where the cavity can quench due to e.g. degradation which lowers quench limits.

EuXFEL SRF cavities are operating in pulsed mode, therefore we have available a sequence of waveforms with a fixed length. The current quench detection system at EuXFEL is observing the quality factor [1]. It is one of the classic methods for the detection and prevention of such quenches. In [2] an online approach for quench detection based on the calculation of detuning and bandwidth on superconducting cavities is presented which is specially tailored to continuous wave operation. A model-based approach for anomaly detection on cavities is shown in [3]. In [4, 5] the parity space method is used to detect anomalies on SRF cavities. Recently data-driven machine learning approaches [6, 7] are used for cavity breakdown prediction on cavities.

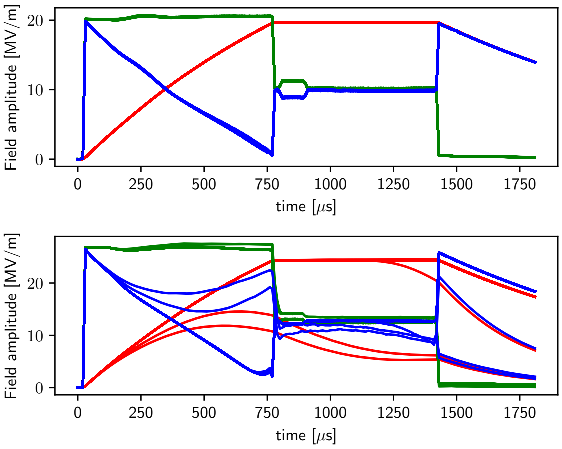

Figure 1: Two examples of amplitudes of healthy (top) and quenching amplitudes (bottom). Probe **p**, forward **f** and **r** signals are depicted red, green, and blue respectively.

Our currently deployed quench detection server [1] provides a daily overview of probable quenches. Recently, EuXFEL created a dataset of events that are probable faults. The availability of such a labeled dataset allowed us to experiment with data-driven machine learning models. This paper presents a study of data-driven anomaly detection to detect faults on RF cavities tailored to the case of EuXFEL. We demonstrate that vanilla data-driven machine learning methods can be trained to predict potential failures with very limited access to training data labeled as faulty.

The structure of this paper is the following: First, we describe the procedure of preprocessing data. Then, we present details of the proposed architecture used for the prediction of faults. Lastly, we show the results of our approach on a test set using different data-driven approaches.

METHOD

Notation

At a time moment t we observe a pulse which consists of three types of complex-valued waveforms: probe **p**, forward **f**, reflected **r**, see Fig. 1. Each event consists of a series of

[*] antonin.sulc@desy.de

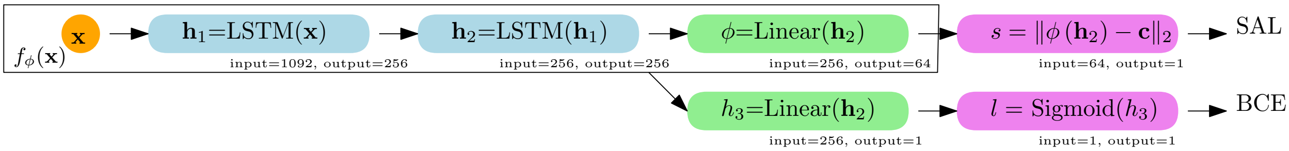
x
$f_\phi(\mathbf{x})$
$\mathbf{h}_1$=LSTM($\mathbf{x}$)
input=1092, output=256
$\mathbf{h}_2$=LSTM($\mathbf{h}_1$)
input=256, output=256
ϕ=Linear($\mathbf{h}_2$)
input=256, output=64
$s = \|\phi(\mathbf{h}_2) - \mathbf{c}\|_2$
input=64, output=1
SAL
h_3=Linear($\mathbf{h}_2$)
input=256, output=1
l = Sigmoid(h_3)
input=1, output=1
BCE

Figure 2: The architecture consists of two LSTM layers. The final linear (green) layer has either 64-dimensional output for a model trained with SAL or 1-dimensional for a model trained with BCE loss. As an input, the network receives a series of signals where a pulse is a 1092 dimensional vector $\mathbf{x}$. For the SAL, the anomaly score s is obtained by calculation of the L_2 distance of the output of the feature layer $f_\theta(\mathbf{x})$ from $\mathbf{c}$. For the model trained with BCE, the likelihood l of a healthy signal is obtained after the application of the Softmax function.

pulses, where all waveforms are 1.82 ms long and sampled at 1 MHz. This yields 1820 values per waveform.

Preprocessing

Processing the entire 1820 values of each waveform is unnecessarily redundant and computationally and data-intensive, therefore we further subsampled each waveform to 182 values. Furthermore, for ease of data handling, we further transformed each type of waveform from amplitude and phase to the IQ coordinates. In summary, we stack a series of transformed IQ waveforms, i.e. 1092 values per pulse $\mathbf{x}$. We further perform normalization of each waveform independently to the $(0, 1)$ range.

The Model

We experiment with two models. One is based on semi-supervised anomaly loss (SAL) the other with binary cross-entropy loss (BCE). For both models, we use identical architecture with recurrent units, but the final layers are slightly different. The proposed model consists of stacked Long Short-Term Memory (LSTM) units with a linear unit, see Fig. 2. The first LSTM layer encodes the input $\mathbf{x}$ into 256-dimensional vector $\mathbf{h}_1$, which is further passed another LSTM layer $\mathbf{h}_2$ with identical dimensionality as $\mathbf{h}_1$. The choice of architecture is not arbitrary. Most of the available training data is recorded with less than 250 pulses. Therefore to cover the entire time range of the labeled faults, we designed a two-layer network where input and hidden recurrent neural units have 256 hidden units. This should provide sufficient freedom to train the temporal relations that are contained in individual training events.

Semi-supervised Anomaly Detection Values from $\mathbf{h}_2$ are presented into the final linear layer that produces a 64-dimensional vector ϕ. We refer to transformed inputs into the final linear layer as features $f_\theta(\mathbf{x})$. Calculation of anomaly score s is performed by measuring the L_2 distance of the input's features $f_\theta(\mathbf{x})$ from a common centre $\mathbf{c}$.

Classifier Unlike the model trained with SAL, the classifier trained with BCE has a single binary value that signifies if the output is faulty or not, therefore the last vector $\mathbf{h}_2$ is fed into a linear unit with just one output h_3. A sigmoid unit can be used to obtain a likelihood of a healthy signal l.

Optimization

The critical component of our model is a proper loss function. Since we have only a few labeled training data as a fault, the BCE may suffer from biases toward healthy data. For this purpose, we adopted SAL [8] defined over N training samples as

$$L(\theta) = \frac{1}{N}\sum_{i}^{N} \|f_\theta(\mathbf{x}_i) - \mathbf{c}\|_2 + \eta \|f_\theta(\mathbf{x}_i) - \mathbf{c}\|_2^{y_i}. \quad (1)$$

The first term is a one-class loss [9] that fits the input data regardless of label. The second term is the SAL [8] which takes into account the category of the input data $\mathbf{x}$ by exponentiating by y. Variable y signifies if $\mathbf{x}$ is an anomaly or not. Intuitively, if y is not an anomaly, then the value is set to 1 and the optimization trains the network and $\mathbf{c}$ to move as close as possible. Contrarily, if y is a (known) anomaly, the value is set to -1 and it intuitively moves the network $f_\theta(\mathbf{x})$ and $\mathbf{c}$ away from each other thus increases the anomaly score for such samples. Since we have a small set of partially labeled data that contains various types of faults, the trained model mostly adapts to standard operations in healthy data with a noticeable emphasis on unhealthy data which are moved away from each other. The benefit of SAL is that features $f_\theta(\mathbf{x})$ reveal information about various faults in the feature space since it optimizes distances in it. It potentially allows further analysis with very little supervision because the network itself plays a role of a bottleneck. It is important to note that unlike [8], we update $\mathbf{c}$ during optimization.

We also evaluated the LSTM model with BCE loss. It requires minor changes in architecture and the output of $\mathbf{h}_2$ is replaced with a single-valued binary output h_3. After applying sigmoid, we have a likelihood l of a healthy pulse. This architecture has an important merit because evaluation provides a likelihood of whether the input is an anomaly or not and avoids the need to specify a threshold.

Each event consists of a sequence K of pulses $(\mathbf{x}_1, \dots \mathbf{x}_K)$. For healthy events, the loss is calculated for all pulses in the event, in case of faulty events only the last event is trained.

EXPERIMENTS

Faulty data was detected by an available quench detection server [10]. Faulty events consist of 250 pulses or less with usually 200 pulses before the event and 50 afterward.

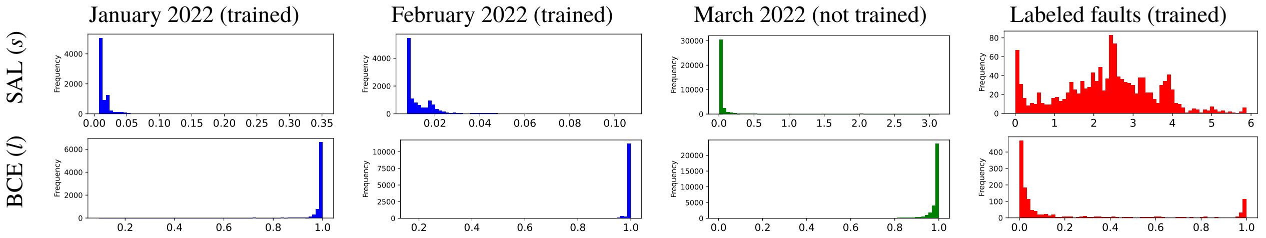

Figure 3: Histograms of anomaly scores s trained with SAL (Top) and likelihoods l trained with BCE (Bottom) for different datasets. (Top) The events sampled from trained months (Jan, Feb) have very low anomaly score s and the highest scores are always below 0.35. Events sampled in March have an increase in anomaly scores since there are seasonality effects that might be labeled as an anomaly. (Bottom) Likelihood of a healthy last pulse l of a model trained with BCE loss.

Table 1: Evaluation of model trained models on two test sets. The test set consists of samples randomly sampled from available data. March 2022 is sampled over the entire month to test how the model responds to long-term untrained events.

	Test set				March 2022	
Method	**TP**	**TN**	**FP**	**FN**	**TN**	**FP**
SAL	**103**	**7691**	**0**	**9**	**35034**	**952**
BCE	96	7685	6	16	34869	1117

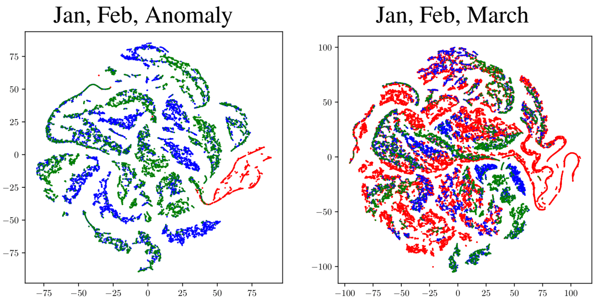

Figure 4: T-SNE Embedding [11] of feature outputs $f_\theta(\mathbf{x})$ of SAL on the last pulse in the event. (Left) Training and testing events from Jan 2022 (green), Feb 2022 (blue) and anomaly (red). (Right) Training and testing events from Jan. 2022 (green), Feb 2022 (blue) and March 2022 (red).

The healthy data were equally sampled and downloaded from our DAQ system [12]. Healthy events usually have 250 or 500 pulses. The training data are sparsely sampled from all available stations over a period of five months.

We have 81922 healthy events and 1331 labeled faulty events available. The test set is randomly sampled with 7803 healthy and 102 faulty events. Additionally, we created one test set with only healthy events over a period of March 2022 to test how models react to novelty.

Evaluation

We evaluated both approaches after 50 epochs trained with ADAM. Table 1 and Fig. 3 show their comparisons.

The model trained with SAL performs better on the test set, where none of the healthy events was wrongly identified (FP) and only 9 faulty events were not identified (FN), see Table 1. The threshold for labeling an event as an anomaly was all scores exceeding one times the standard deviation of all scores in the test set. Slightly better performance than BCE is noticeable on the healthy test set sampled in March 2022, where 952 events were identified as a fault. This can further be identified in the right image in Fig. 4, where a part of the healthy events of March 2022 noticeably deviates from the trained datasets for January and February 2022.

The model trained with BCE wrongly identified 6 healthy events (FP) as faulty and 16 faulty events were not identified (FN). The model also performs slightly worse on the March 2022 test set by identifying 1117 healthy events as faulty.

CONCLUSION AND FUTURE WORK

In this paper, a data-driven and model-free approach to detecting cavity anomalies is shown. We show a training model which can bypass the disproportionally many healthy training samples by using a SAL [8] and compare it with BCE. The SAL allowed us to train the proposed model with an abundance of healthy data. As a byproduct, the model is trained to project inputs to a feature space that reveals the potential for further classification of different types of faults.

Experiments show that our method can identify a large part of faults in our test set. One of the major limitations is that waveforms may vary over longer periods. This causes a noticeable increase in false positives for events from different time periods.

In the future, since the lower-dimensional features of the SAL model still carry the information about a fault, we would like to experiment with different models to achieve better interpretability of the features and distinguish between different types of faults. Since the dimensionality of features is much smaller, this should also require smaller training datasets. Additional insights can also be revealed by using generative models for anomaly detection [13] instead of discriminative ones.

ACKNOWLEDGEMENT

We thank Julien Branlard for his helpful suggestions and corrections on the paper.

REFERENCES

[1] J. Branlard, V. Ayvazyan, O. Hensler, H. Schlarb, Ch. Schmidt, and W. Cichalewski, “Superconducting Cavity Quench Detection and Prevention for the European XFEL”, in *Proc. ICALEPCS'13*, San Francisco, CA, USA, Oct. 2013, paper THPPC072, pp. 1239–1241.

[2] A. Bellandi and *et al.*, “Online detuning computation and quench detection for superconducting resonators,” *IEEE Transactions on Nuclear Science*, vol. 68, no. 4, pp. 385–393, 2021. doi:10.1109/TNS.2021.3067598

[3] A. S. Nawaz, S. Pfeiffer, G. Lichtenberg, and P. Rostalski, “Anomaly Detection for Cavity Signals - Results from the European XFEL”, in *Proc. IPAC'18*, Vancouver, Canada, Apr.-May 2018, pp. 2502–2504. doi:10.18429/JACoW-IPAC2018-WEPMF058

[4] A. S. Nawaz, S. Pfeiffer, G. Lichtenberg, and P. Rostalski,, “Anomaly detection for the european xfel using a nonlinear parity space method,” *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 1379–1386, 2018. doi:10.1016/j.ifacol.2018.09.554

[5] A. Eichler, J. Branlard, and J. H. K. Timm, “Anomaly detection at the european xfel using a parity space based method,” doi:10.48550/arXiv.2202.02051, 2022.

[6] C. Tennant, A. Carpenter, T. Powers, A. S. Shabalina, L. Vidyaratne, and K. Iftekharuddin, “Superconducting radio-frequency cavity fault classification using machine learning at Jefferson laboratory,” *Physical Review Accelerators and Beams*, vol. 23, no. 11, p. 114601, 2020. doi:10.1103/PhysRevAccelBeams.23.114601

[7] C. Obermair *et al.*, “Machine Learning Models for Breakdown Prediction in RF Cavities for Accelerators”, in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 1068–1071. doi:10.18429/JACoW-IPAC2021-MOPAB344

[8] L. Ruff, R. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K. Müller, and M. Kloft, “Deep semi-supervised anomaly detection,” doi:10.48550/arXiv.1906.02694, 2019.

[9] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *International conference on machine learning*, PMLR, 2018, pp. 4393–4402.

[10] J. H. K. Timm, J. Branlard, A. Eichler, and H. Schlarb, “The Trip Event Logger for Online Fault Diagnosis at the European XFEL”, in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 3344–3346. doi:10.18429/JACoW-IPAC2021-WEPAB293

[11] G. Hinton and S. Roweis, “Stochastic neighbor embedding,” *Advances in neural information processing systems*, vol. 15, 2002.

[12] T. Wilksen *et al.*, “The Control System for the Linear Accelerator at the European XFEL: Status and First Experiences”, in *Proc. ICALEPCS'17*, Barcelona, Spain, Oct. 2017, pp. 1–5. doi:10.18429/JACoW-ICALEPCS2017-MOAPL01

[13] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157.