Christoph Obermair

Interpretable Fault Prediction for CERN Energy Frontier Colliders

DOCTORAL THESIS

to achieve the university degree of
Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

TU-Graz Supervisor

Franz Pernkopf

Signal processing and speech communication laboratory

CERN Supervisor

Daniel Wollmann

Controls and beam studies for protection section

Graz, Jan. 2024

CERN-THESIS-2023-321
19/01/2024

Abstract

The Large Hadron Collider (LHC) is the world's highest energy particle collider, which has already delivered data for numerous physical discoveries. To continue this quest for discovering new physics, the Compact Linear Collider (CLIC) and the Future Circular Collider (FCC) aim to push the boundaries of fundamental physics at high collision energies. However, as their power, size, and complexity increases, so does the risk of failures and their associated downtime.

Fault prediction is a way to minimize downtime by fixing faults in scheduled maintenance intervals before they occur. In the LHC, such fault prediction methods have been supporting system experts to decrease downtime since its start in 2008/9. There are many different scenarios of faults. Each of them occurs rarely, which is why the predictions cannot be validated by statistical tests alone. Nonetheless, the methods work reliably as their predictions are based on known fault indicators which are validated by experts. To use Machine Learning (ML) methods for fault prediction, the same approach is required: Predictions must be interpreted and validated by experts.

Demonstrating the predictive capabilities of ML, this thesis presents three approaches for interpretable fault prediction. Firstly, a novel autoencoder-based method for explaining fault predictions to system experts is proposed. A survey of 73 potential users confirms its effectiveness when compared to two other popular methods.

This explanation method is then used to interpret ML-based breakdown predictions in radio frequency cavities. The interpretation reveals that a pattern in the emitted electrons following an initial breakdown is closely related to the probability of another breakdown occurring shortly thereafter. This explanation is consistent with the findings of recent research.

Secondly, non-negative matrix factorization, a ML method that is designed to be interpretable, is used to detect normal and abnormal behavior in the LHC main dipole magnets. Five dipole magnets with abnormal behavior are identified, of which one was confirmed to be damaged.

Thirdly, a hybrid method is proposed, that allows experts to rely on their existing tools and still benefit from non-interpretable ML. The method is tested to predict faults in a protection system of the LHC main dipole magnets. The method captured 113 out of 116 faults, while only 99 out of 116 faults were captured with the existing tool.

The presented approaches demonstrate the strength of interpretable ML for contributing to reliable operation of next-generation particle accelerators. Its applicability extends to numerous other collider components, including radio frequency cavities and dipole magnets in the FCC.

List of Publications

This thesis includes content from the following publications:

1. C. Obermair, M. Maciejewski, F. Pernkopf, Z. Charifoulline, A. Apollonio, and A. Verweij, "Machine Learning with a Hybrid Model for Monitoring of the Protection Systems of the LHC," in *Proc. IPAC'21*, JACoW, 2021, pp. 1072-1075. Available:

   https://accelconf.web.cern.ch/ipac2021/papers/mopab345.pdf.

2. C. Obermair et al., "Machine Learning Models for Breakdown Prediction in RF Cavities for Accelerators," in *Proc. IPAC'21*, JACoW, 2021, pp. 1068-1071. Available:

   https://jacow.org/ipac2021/papers/mopab344.pdf.

3. H.S. Bovbjerg, C. Obermair, A. Apollonio, T. Cartier-Michaud, W. Millar, Z.H. Tan, M. Shen, D. Wollmann, "Data Augmentation for Breakdown Prediction in CLIC RF Cavities," in *Proc. IPAC'22*, JACoW, 2022, pp. 1553-1556. Available:

   https://accelconf.web.cern.ch/ipac2022/papers/tupoms054.pdf.

4. C. Obermair, T. Cartier-Michaud, A. Apollonio, W. Millar, L. Felsberger, L. Fischl, H. S. Bovbjerg, D. Wollmann, W. Wuensch, N. Catalan-Lasheras, M. Boronat, F. Pernkopf, G. Burt, "Explainable Machine Learning for Breakdown Prediction in High Gradient RF Cavities," *Phys. Rev. Accel. Beams*, APS, vol. 25, no. 10, p. 104601, 2022. Available:

   https://link.aps.org/doi/10.1103/PhysRevAccelBeams.25.104601.

5. C. Obermair, A. Fuchs, F. Pernkopf, L. Felsberger, A. Apollonio, D. Wollmann "Example or Prototype? Learning Concept-Based Explanations in Time-Series," in *Proc. ACML'14*, PMLR, 2023, pp. 816-831. Available:

   https://proceedings.mlr.press/v189/obermair23a/obermair23a.pdf.

6. C. Obermair, A. Apollonio, Z. Charifoulline, L. Felsberger, M. Janitschke, F. Pernkopf, E. Ravaioli, D. Wollmann, M. Wozniak, "Interpretable Anomaly Detection in the LHC Main Dipole Circuits with Non-negative Matrix Factorization," submitted to IEEE Transactions on Applied Superconductivity, 2023

Acknowledgement

I am deeply grateful for the exceptional opportunity to undertake this work, a journey made possible by the invaluable support and contributions of many individuals.

In particular, I would like to thank Franz Pernkopf, my supervisor of the Technical University of Graz. I always felt welcome in his Intelligent Systems group, and enjoyed the best possible technical and organizational support. Having spent most of my time at CERN in Geneva, this cannot be taken for granted.

I would like to thank Andrea Apollonio, who supervised me during my first two years at CERN. He made it easy for me to get started on my Ph.D. topic by looking ahead and making sure I had access to all the necessary resources before I was even aware of their importance.

I deeply express my gratitude to Daniel Wollmann, our section leader and my CERN supervisor during the last year of my Ph.D. I am thankful for his flexibility which allowed me to work at the TU Graz when required. Despite his tight schedule, he always found time for a weekly meeting with me. Besides the technical aspects, he taught me a lot about the importance of consistency, leadership, and diplomacy.

I am grateful to Rüdiger Schmidt for his prompt availability to review my thesis and his detailed feedback.

My work certainly would not have been possible without Lukas Felsberger, Alexander Fuchs, and Thomas Cartier-Michaud. As unofficial supervisors, they consistently demonstrated remarkable dedication, never hesitating to find coding bugs, proofreading texts, and engaging hands-on in various aspects of my work.

Working together with my colleagues in the CB section and at the SPSC Institute made not only work exciting and entertaining, but also the lunch and coffee breaks. Special thanks to my friends, particularly the Schuastahütte, the Wist-shiners, and the Klaus-mates, for always being there for a drink, even when my time was scarce. Together with Lukas, I could ensure the constant availability of these electrolyte-containing drinks.

I am deeply grateful for the privilege of having such supportive parents. Their encouragement, guidance, and trust, have been my pillars of strength and inspiration along my educational journey.

Finally, I express my sincere gratitude to my beloved girlfriend Melanie. Our weekly Zoom dates, filled with cheese-making, artistry, cooking, and much more, played an essential role in keeping me grounded and focused on the important aspects of life. I am thankful for her patience, her unconditional support, and above all, for her consistently positive attitude.

List of Acronyms and Abbreviations

ADC Analog to Digital Converter. 38

AE Autoencoder. 25–27, 31, 32

AR Area under the Receiver operating characteristics curve. 39

CB Crowbar. 44

CL Current Leads. 44

CLIC Compact Linear Collider. 3–8, 10, 12–18, 25, 31, 37, 38, 55

DBSCAN Density-Based Spatial Clustering of Applications with Noise. 24

DL Deep Learning. 21, 39

EBE Explanation-By-Example. 31, 33, 35

EE Energy Extraction. 45

FC Faraday Cups. 38

FCC Future Circular Collider. 3, 4, 7, 8, 10–14

FFT Fast Fourier Transformation. 47, 49

FP False Positive. 53

FPA Fast Power Abort. 44–50, 56

FPM Frequency Position Map. 47, 48

HF High Field. 45

HWC Hardware Commissioning. 45

LF Low Field. 45, 51, 52

LHC Large Hadron Collider. 3–18, 20, 21, 23, 24, 43, 45, 50, 53, 55, 56

LS Long Shutdown. 9

MAP Model-Agnostic Prototype. 32, 33, 35

ML Machine Learning. 4–6, 19, 21, 29, 34, 35, 39, 43, 46, 52, 55, 56

MSP Model-Specific Prototype. 31, 33, 35

NMF Non-negative Matrix Factorization. 23–27, 47, 56

NN Neural Network. 21, 22, 25, 27, 31

PC Power Converter. 44, 45

PCA Principal Component Analysis. 23, 24

QDS Quench Detection System. 44

QH Quench Heater. 6, 44–46, 49, 51–53

QHDA Quench Heater Discharge Analysis. 6, 51–53, 55

ReLU Rectified Linear Unit. 21, 22

RF Radio Frequency. 4–7, 9–16, 18, 20, 23, 24, 31, 37–40, 55

SPS Super Proton Synchrotron. 14

SSPA Solid State Power Amplifier. 14

STEAM Simulation of Transient Effects in Accelerator Magnets. 15

SVM Support Vector Machine. 20–22, 27, 30, 52, 53, 55

t-SNE t-distributed Stochastic Neighbour Embedding. 23, 24

TP True Positive. 53

XAI Explainable Artificial Intelligence. 4, 5, 29, 32, 35, 40, 41, 55

Contents

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. I have used language models strictly for editorial assistance, such as grammar and style refinement, and not for the generation of original content or ideas central to this research. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date	Signature

Part I

Overview and Contributions

Chapter 1

Introduction

1.1 Energy Frontier Colliders: Opportunities and Challenges

Accelerators for particle physics are designed for the validation and extension of physical models by analyzing particle collisions. These colliders aim to push the boundaries of higher collision energy, necessary for this purpose.

Accelerators for particle physics are designed for the validation and extension of physical models by analyzing particle collisions. Among accelerators for particle physics, colliders aim to push the boundaries of higher collision energy, necessary for this purpose.

The Large Hadron Collider (LHC) is the world's current energy frontier collider, colliding two proton beams with an energy of 13.6 TeV [1]. The energy stored in both beams is equivalent to that of a 400-ton TGV train traveling at 150 km/h [2]. Possible next-generation energy frontier colliders at CERN, the European Organization for Nuclear Research, are the Compact Linear Collider (CLIC), and the Future Circular Collider (FCC). CLIC will collide electrons and positrons with collision energies up to 3 TeV [3]. FCC will be built in two steps: Initially aiming for electron-position collisions with up to 350 GeV [4], followed by proton-proton collisions up to 100 TeV [5].

To validate and extend the physical models thoroughly, not only a high energy is required, but also a high integrated luminosity. This term is equivalent to the total number of collisions accumulated over time in a particle accelerator. The higher the integrated luminosity, the more statistics are available to validate and extend the physical models. In CERN's next-generation colliders, a crucial factor to reach high integrated luminosity is the time the accelerator is fully operational and available for colliding particle beams. This time is referred to as availability. In case the machine has to be switched off during operation due to a hardware fault, the downtime for repairs negatively impacts the availability. The downtime should be reduced to a minimum, to reach high integrated luminosity.

Fault prediction enables fault prevention, by repairing the faulty components in scheduled maintenance intervals. This reduces machine downtime during operation and increases machine availability.

One challenge for fault prediction methods is that there are many different scenarios of faults, where each of them only occurs rarely. This leads to an

insufficient number of data to test the methods extensively and ensure their reliability. Nonetheless, existing fault prediction methods used at CERN [6], [7] still work reliable. This is because they monitor known fault indicators, validated by system experts.

In recent projects, also the ability of Machine Learning (ML) to detect non-linear relationships in large amount of data has been useful for fault prediction [8], [9]. At CERN, more and more data is being recorded, making the use of ML-based fault prediction increasingly promising. Nonetheless, to ensure that faults can still be predicted reliable, ML is required to be interpretable. Then it is possible to find the fault indicator on the basis of which the prediction is made, and validate that it is a true fault precursor and not a bias.

This thesis investigates methods for interpretable fault prediction in energy frontier colliders with ML. The effectiveness of the methods is demonstrated by two case studies on predicting faults in two cruicial collider components: the CLIC Radio Frequency (RF) cavities and the LHC main dipole magnets. The applicability of the methods extends to numerous components of next-generation particle colliders, including RF cavities and dipole magnets in FCC.

1.2 Research Questions and Contributions

In the following, three research questions (RQs) are presented, which will be answered as part of the contribution of this thesis. While the first research question addresses limitations for interpreting ML-based fault predictions, the other two focus on challenges for predicting faults in CLIC RF cavities and LHC main dipole magnets.

RQ1: How can system experts best interpret machine learning models to obtain reliable fault predictions?

The success of machine learning models frequently relies on the use of complex non-linear functions. These functions make it difficult to identify the indicator on the basis of which the prediction is made [10], [11]. Therefore, system experts cannot determine whether the prediction is reliable or based on a bias. One tool for explaining the ML predictions to humans is Explainable Artificial Intelligence (XAI). This tool can also be used for fault prediction. Studies show, however, that explanations from different XAI methods are perceived differently based on the domain they are used in [12]. Since existing XAI tools have not been tested for fault prediction, the question is which XAI methods to use and whether existing approaches can be further improved.

Research Contribution: Within this thesis, a novel XAI method for explaining fault predictions to system experts is proposed. To evaluate the effectiveness of the method, 73 individuals from CERN and TU-Graz were asked to identify faults in data, using the fault explanations from the proposed method and two other XAI methods. With the proposed method, people were able to predict 79.3% of the faults, while with the other two methods only 73.8% and 67.2% of faults were predicted. The explanations are therefore understood 5.5% and 12.1% better. For CERN system experts, the customized explanations will help them better understand and validate the results of ML. This paves the

way for employing ML-based fault prediction, which will enhance the rate of predicted faults over time.

In the following use cases for predicting faults in CLIC RF cavities and in LHC superconducting dipole magnets, the focus on interpreting ML-based fault predictions remains. Among other methods, the proposed XAI method is used to make fault predictions for energy frontier colliders more reliable.

RQ2: Can data measured at CLIC RF cavities provide insights on breakdown prediction?

RF cavities accelerate the particles using an oscillating electric field. A major constraint to reaching a high accelerating gradient are RF cavity breakdowns. During a breakdown, deformations or contaminations on the cavity surface can cause local field enhancement, leading to electrical arcs and damage to the cavity. Prior to the arc formation, electrons are emitted, which are measured by sensors on the outside of the RF cavities. In a test stand of CLIC RF cavities, the maximal number of emitted electrons is monitored by a threshold to detect breakdowns. The question is if the temporal change of emissions before a breakdown could be used to predicted them at an earlier stage. The data gathered at the CLIC RF cavity test stand, could provide information about this.

Research Contribution: With the data gathered at the CLIC RF cavity test stand, breakdowns are predicted in the preceding pulse. Depending on the type of breakdown, an accuracy of up to 89.7% is achieved. This means that in 89.7% of these cases, the negative consequences of a breakdown could have been mitigated, by suitable adjustment of the electromagnetic field in the RF cavity. The interpretation of the ML models with XAI shows that a pattern in the emitted electrons following an initial breakdown is closely related to the probability of another breakdown occurring shortly thereafter. These observations are consistent with recent research on the emergence of breakdowns [13], [14].

RQ3: How can the data measured at the LHC main dipole magnets provide insights about their normal and abnormal behavior?

The LHC main dipole magnets bend the charged particles with a magnetic field along the trajectory of the accelerator. To reach the nominal field of 8.0 T and a current of 11.85 kA, each magnet is cooled down to 1.9 K with superfluid helium. A fault in a cooled down component, can lead to a downtime of up to three months, as the components have to be warmed up for repair and subsequently cooled down again in a sophisticated procedure. In this regard, there exist several threshold based fault prediction methods [15] and physical models that accurately simulate the circuit behavior of the main dipoles [16]. With the existing data of the main dipole magnets, ML could provide additional insights into their normal and abnormal behavior. These insights would further improve existing fault prediction methods and physical models.

Research Contribution: Within this thesis, normal and abnormal frequency patterns are identified in the voltage measured at the dipole magnets. The characteristics and possible origins of these patterns are analyzed, to

improve the existing physical models. The patterns measured at four dipole magnets are identified to be an indicator of a fault. One of these magnets already failed, providing information about the method's reliability. The results lead to the scheduling of additional measurements. If a fault cannot be excluded during these measurements, the magnets could be replaced in one of the next maintenance stops of the LHC.

Furthermore, faults are predicted in the resistive strips inside the magnets. These Quench Heaters (QHs) protect the superconducting magnet coils in case of a resistive transition, also called quench. An existing Quench Heater Discharge Analysis (QHDA) tool already ensures that faults in the QHs do not cause a fault in the magnet. This tool is extended with novel hybrid method. The hybrid method allows experts to rely on their existing tool and benefit from fault predictions with ML. In a test, this method is able to capture 113 out of 116 faults, while only 99 out of 116 faults are captured with the existing QHDA tool. In the tree-year period in which the test data was recorded, therefore, the risk of a fault in a magnet could have been reduced in 12.3% of all QH faults.

1.3 Outline

The thesis consists of two parts. Part I provides a summary of all publications and states the three main contributions of this thesis. An introduction to accelerators and their challenges is given in Section 2. The machine learning methods, used for fault prediction in the applications of this thesis, are introduced in Section 3. The methods for interpreting ML are described in Section 4. Section 5 summarizes the work of predicting breakdowns in CLIC RF cavities and Section 6 shows the detection of normal and abnormal behavior in the LHC main dipole magnets. Finally, the thesis summarizes the contributions, states open questions, and elaborates on possible future work. In Part II, the publications of the thesis are included.

Chapter 2

Accelerators

Particle accelerators use a beam of charged particles for a variety of applications, some of which are described below. Medical applications use the beam to sterilize medical devices [17], diagnose and destroy cancer cells [18], for particle beam radiography [19], and to produce of Positron Emission Tomography (PET) tracers [20]. In industry, the beam is projected onto a target material to make it more durable [21] or to enhance certain of its properties [22]. Particle accelerators are also used in security applications to scan cargo [23] or for environmental research to study pollutants [24]. For particle physics, the beam is accelerated and brought to collision with a fixed target or with another counter-rotating beam to experimentally validate and extend physical models.

This chapter introduces three energy frontier colliders for particle physics at CERN: LHC, CLIC, and FCC. It is structured as follows: Section 2.1 gives an overview of the history and functionality of the colliders. The number of their RF cavities and superconducting dipole magnets are discussed to emphasize the scale of these projects. Section 2.2 presents a literature review of existing availability studies. These studies show, that additional innovations are necessary to meet the availability targets of CLIC, and FCC. Fault prediction is a compelling solution.

2.1 Energy Frontier Colliders at CERN

CERN is the European Organization for Nuclear Research with currently 23 member states, located in the west of Switzerland at the border to France. It provides essential infrastructure and equipment for fundamental research in particle physics to over 12200 scientists of 110 nationalities [25]. At the heart of CERN lies the LHC, where the discovery of the Higgs boson was achieved in 2012 [26]. To further contribute to the scientific understanding, the European Strategy for Particle Physics proposes the construction of particle colliders with a larger energy reach [27]. Therefore, the feasibility of two candidates for next-generation accelerators is investigated: CLIC and FCC. A map with the planned locations and the sizes of these accelerators are shown, together with the existing LHC, in Fig. 2.1.

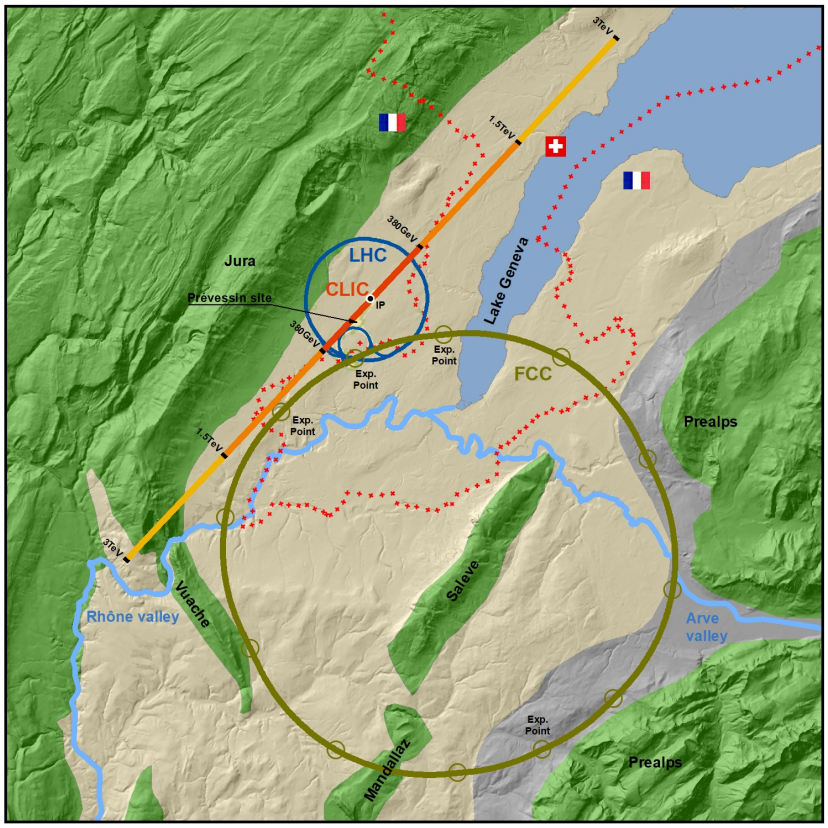


Figure 2.1: The geographical location and size of the LHC, the CLIC and the FCC [28]. While the LHC is in operation since 2008. CLIC, and FCC are potential next-generation colliders.

2.1.1 Large Hadron Collider (LHC)

In the early 1980s, it was proposed to build a particle accelerator with sufficient beam energy to discover the Higgs boson. The project was approved in December 1994 and the construction of the LHC began in October 1995. Upon completion in September 2008, the LHC faced one year of downtime, due to a fault in a main superconducting dipole circuit [29]. The first operational run began in November 2009 and continued successfully until February 2013. After that, technical upgrades and maintenance work were carried out within the first Long Shutdown (LS), which lasted until 2015 [30]. More upgrades and maintenance work took place during LS2 from 2019-2021 [31] and are planned for LS3 in 2026-2028 [32]. In LS3, the High Luminosity LHC (HL-LHC) upgrade will significantly improve the technical performance [33], [34].

The LHC consists of a 27 km long ring in which two proton or heavy ion beams circulate on opposite trajectories. Current beam energy can be up to 6.8 TeV, which corresponds to a velocity close to the speed of light. To obtain this velocity, the beams are first pre-accelerated trough various injectors. Each injector contributes increasingly more energy to the beam until it is injected into the LHC at an energy of 450 GeV. CERN's accelerator and experimental complex is illustrated in Fig. 2.2. The yellow dots represent the four LHC interaction regions at which the two beams collide. At these points, the particles created during the collisions are analyzed by highly specialized detector arrangements [35].

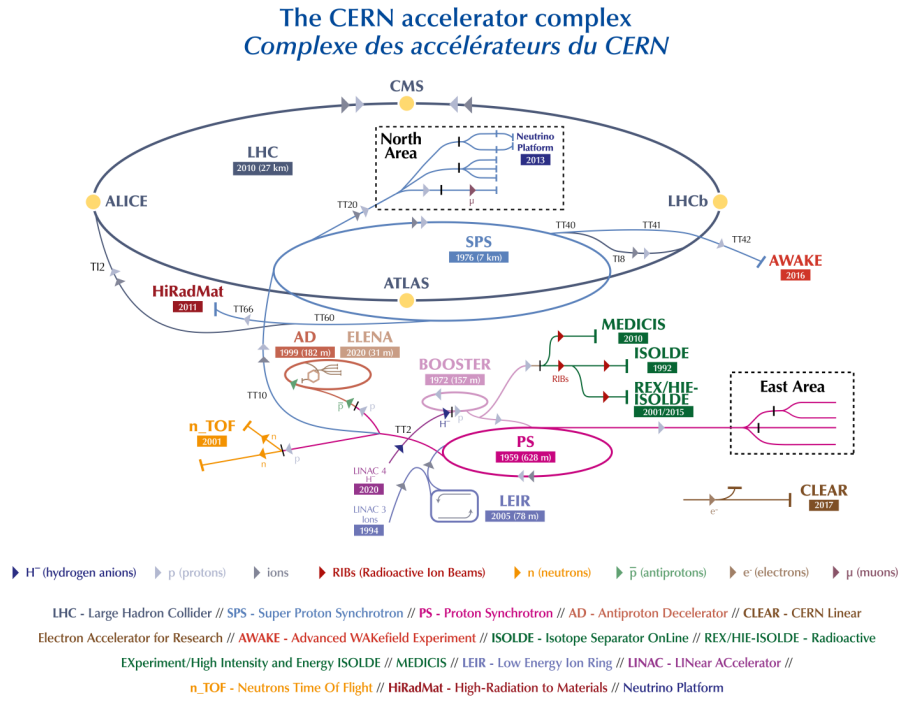


Figure 2.2: Overview of the CERN accelerator complex [36].

The beams are accelerated by an oscillating electric field in the RF cavities

and steered along their circular path by magnetic fields from magnets. This can be expressed in terms of the Lorenz force [37]

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}), \tag{2.1}$$

where q is the electrical charge of the particles, $\vec{E}$ is the electric field, $\vec{v}$ is the velocity of the particles, and $\vec{B}$ is the magnetic field. In the LHC, 1232 superconducting dipole magnets provide the bending force, each with a nominal magnetic field of 8.3 T [1]. To reach this field, each magnet is cooled down to 1.9 K with superfluid helium.

The accelerating force is given by eight superconducting RF cavities per beam. Each cavity provides an 400 MHz accelerating gradient of 5.3 MV/m [38], defined as the maximum voltage gained by a relativistic particle while passing the cavity, divided by the length of the accelerating gap [39]. With this gradient, protons are currently accelerated from 450 GeV to 6.8 TeV within a 20 minutes period. This period is referred to as *ramp-up*, during which the magnetic field of the main dipole magnets is gradually increased with the beam intensity. +

2.1.2 Compact Linear Collider (CLIC)

CLIC is one of the candidates for the next generation of CERN's energy frontier colliders. The proposal to implement the electron-positron collider was published in December 2018 [3]. The implementation is foreseen in three stages, in which the accelerator is continuously being extended from a length of 11 km to 29 km and 50 km to reach a collision energy of 380 GeV, 1.5 TeV and 3 TeV, respectively. In the project implementation plan, it is mentioned that collisions could begin by 2035 if the project is approved and launched by 2026 [3, p. 193]. The data from these collisions would permit the analysis of the Higgs boson's interactions with other particles and with itself. The precision of these measurements would be significantly higher than those of proton–proton collisions in the LHC [40]. Each of the upgrades to the 1.5 TeV and 3 TeV stage would take approximately 4 years.

Particles in CLIC are accelerated along a linear trajectory, without the need for bending magnets. Due to this linear trajectory, particles can only be accelerated once by each RF cavity. Already for the 380 GeV stage 21630 RF cavities would be required [41]. These cavities would be normal conducting and provide a pulsed accelerating gradient of 70 to 100 MV/m oscillating at 12 GHz [42].

2.1.3 Future Circular Collider (FCC)

Another candidate for a next-generation collider is the FCC. The conceptual design report was published in 2018, stating the intention to study Higgs boson's interactions, dark matter, antimatter, and the overall understanding of universe's fundamental forces and constituents [43]. As part of the FCC project, a ∼91 km long tunnel is equipped with two particle accelerators. Initially an electron-positron collider is build, allowing collision energies larger than 350 GeV (FCC-ee) [4]. Then a proton and heavy ion collider with up to 100 TeV (FCC-hh) [5] is constructed. If the design report is approved by 2030, the construction of the FCC-ee is foreseen to start in the mid 2030s. This would

allow collisions from 2045 to 2060. Construction of the FCC-hh then begins, with collisions foreseen in the 2070s.

Technical Requirements

An electron has ~ 1800 times lower rest mass m_0 than a proton. Therefore, a lower magnetic field is required to overcome the relativistic centrifugal force

$$F_c = \frac{\gamma m_0 v^2}{\rho}, \tag{2.2}$$

where $\gamma = \frac{1}{\sqrt{1-(v^2/c^2)}}$, c is the speed of light, and ρ is the bending radius [44]. In the FCC-ee, 2900 normal conducting dipole magnets with a magnetic field of 14.1–56.6 mT are foreseen [4].

However, when charged particles are bent by magnetic fields at relativistic speeds they emit photons. This synchrotron radiation results in an energy loss of

$$\Delta E = \frac{e^2}{3\epsilon_0 (m_0 c^2)^4} \frac{E^4}{\rho} \tag{2.3}$$

per turn [45]. Here, e is the elementary electric charge, ϵ_0 the vacuum permittivity, and E the energy of the circulating particles. The influence of the mass is $\Delta E \sim m_0^{-4}$, which is why circular electron accelerators require continuously high RF acceleration to compensate for the loss of energy due to synchrotron radiation. Therefore, the FCC-ee will use up to 1352 superconducting RF cavities per beam. Each cavity is providing a continuous wave oscillating gradient of up to 20.1 MV/m at 400-800 MHz, based on the mode of operation [46, slide 20].

The FCC-hh, requires higher bending forces, and potentially 4668 superconducting dipole magnets with a magnetic field strength of up to 16 T. This is accompanied by 24 superconducting RF cavities per beam with an accelerating gradient of 5.3 MV/m oscillating at 400 MHz. These RF cavities are similar to the ones used in the LHC [5]. To serve as an injector chain for the FCC-hh, the existing accelerator complex of the LHC (see Fig. 2.2) will be modified [47].

2.2 Availability Requirements

Availability is a crucial performance measure for accelerator engineering. It refers to the proportion of time in which the collider is ready to deliver beam for physics experiments during planned operation [48]. Planned operation T_{op} is the time at which the machine is expected to be fully functional in accordance with the operational schedule. This time is used to process the beam (e.g. ramp-up) [49] and for collisions with a stable beam. Planned shutdowns and commissioning periods are generally excluded from T_{op}. However, downtime due to unplanned maintenance and repairs negatively affects availability.

The target availability is the minimal availability that is required to achieve the defined physics objective of a particle accelerator. This objective depends on the number of collisions and is typically given as the integrated luminosity L_{int} in inverse femtobarn (fb^{-1}). For particle accelerators with a constant potential

number of collisions per seconds L, the target availability is derived by:

$$A = \frac{L_{\text{int}}}{L \times T_{op}}. \tag{2.4}$$

L is also referred to as the nominal luminosity of the collider and is typically given in $\text{cm}^{-2}\text{s}^{-1}$ ($1\ \text{fb}^{-1}=10^{39}\ \text{cm}^{-2}\text{s}^{-1}$). In this section, the target availability of CLIC, and FCC is assessed, and compared to the observed availability of the LHC from 2016 to 2018.

Fig. 2.3 shows the overall availability of the LHC from 2016 to 2018 for operation with protons. In total, the LHC was ready to deliver beam for physics for 353 out of 459 days, which corresponds to an availability of ∼77%. Excluding the processing of the beam, a stable beam was then delivered on 218 days [50]. The right-hand side of Fig. 2.3 shows how the 106 days of downtime are distributed among the subsystems of LHC. The injector complex caused the highest downtime with 22.2% (23.5 days) followed by the cryogenics with 13.1% (13.9 days) and the electrical network with 10.5% (11.1 days). RF cavities belong to the radio frequency system, which caused 3.8% (4.0 days) of the downtime. Superconducting magnets are part of the magnet circuit system with a downtime of 3.0% (3.2 days).

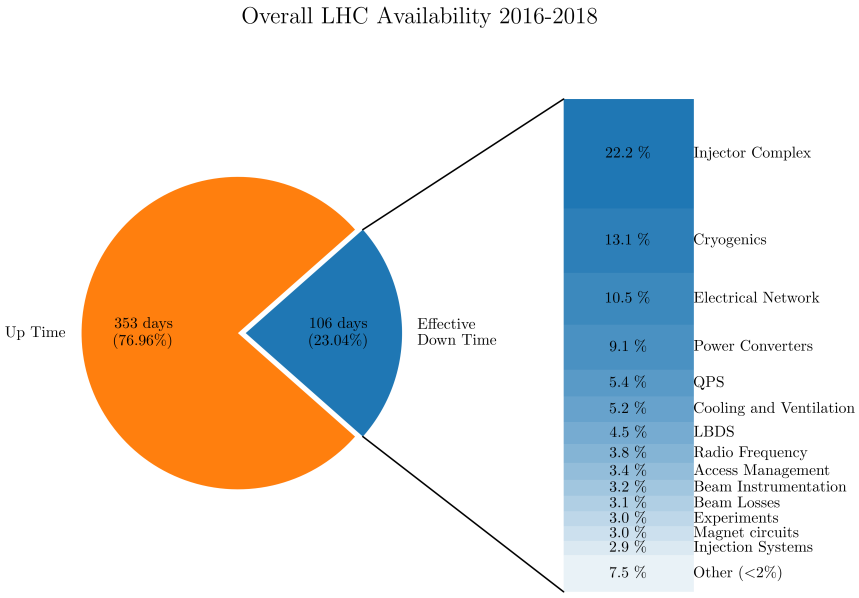


Figure 2.3: Availability of the LHC and its subsystems [51].

Using Eq. (2.4), the availability targets of CLIC, and FCC can be estimated:

1. **CLIC**: In this accelerator 185 days of operation are foreseen per year. Based on the three implementation stages 380 Gev, 1.5 TeV, and 3 TeV, a yearly annual integrated luminosity of 180 fb^{-1}, 444 fb^{-1}, and 720 fb^{-1} at a constant luminosity of 1.5 $\text{cm}^{-2}\text{s}^{-1}$, 3.7 $\text{cm}^{-2}\text{s}^{-1}$, and 6 $\text{cm}^{-2}\text{s}^{-1}$ is planned [52, Tab. 6]. In all three stages, this requires an availability of $A_{\text{CLIC}} = 75\%$.

2. **FCC-ee**: 185 days of FCC-ee operation are dedicated to physics operation at a nominal luminosity of up to $L = 200 \times 10^{34}\ \text{cm}^{-2}\text{s}^{-1}$. The goal is to

Accelerator	Observed/Target Availability	# Main Dipole Magnets	# RF Cavities
LHC	77%	1232*	16*
CLIC	75%	0	21630
FCC-ee	80%	2900	1352*
FCC-hh	70%	4668*	48*

Table 2.1: Overview of observed or target availability in CERN energy frontier colliders. The number of main dipole magnets or RF cavities is marked by a '*' if the components are superconducting. For CLIC the number of RF cavities of the 380 GeV stage are stated.

have an integrated luminosity of up to $L_{\text{int}} = 24$ ab^{-1} ($= 24 \times 10^{42}\text{cm}^{-2}$) per year and detector [4, p. 379]. The luminosity is essentially constant as electrons and positrons are injected "on top" of a circulating beam. Using Eq. (2.4) this requires a target availability of 75%. Due to an additional margin, the design report states a target availability of $A_{\text{FCC-ee}} = 80\%$ [4, p. 377].

3. **FCC-hh**: Per day of operation dedicated to physics, the FCC-hh aims to reach an integrated luminosity of up to $L_{\text{int}} = 8$ fb^{-1}, with a peak luminosity of up to $L = 3 \times 10^{35}$ cm^{-2}s^{-1}. 160 days of operation dedicated to physics are foreseen [52, Tab. 1]. Luminosity in the FCC-hh is however not constant, which is why L has to be integrated over time. This is further described in the design report, showing a target availability of $A_{\text{FCC-hh}} = 70\%$ [5, p. 778].

Table 2.1 summarizes the availability of the LHC from 2016 to 2018 and the target availability of CERN's next generation particle accelerators. In addition, this table shows the number of main dipole magnets and the number of RF cavities. If the components are superconducting, the number is marked with a '*'. Faults in such components, can lead to a downtime of up to three months, as the components have to be warmed up for repair and subsequently cooled down again in a sophisticated procedure.

One can generally assume that downtime increases with the number of components in a system [51]. The order of magnitude more components combined with the ambitious technological objectives means that availability assurance in the future colliders is unlikely to be easier than in the LHC. Taking this into account, the feasibility of the stated target availabilities is discussed in more detail below.

1. **CLIC**: The CLIC implementation plan validates the stated target availability by comparisons to availabilities in the Free Electron Laser (FEL) linear accelerators and light sources of the latest generation [3, p. 15]. CLIC RF cavity breakdowns pose a potential risk to achieve the target availability, as they can lead to spurious interlocks and hardware failures [3, p. 34]. Due to the substantial quantity of cavities (see Tab. 2.1), the impact on the downtime can be significant.

2. **FCC-ee**: Similarly to the LHC and contrary to CLIC, the FCC-ee RF cavities are superconducting. Due to the high number of superconducting RF cavities, a recent study projected the LHC RF availability onto

the FCC-ee case, assuming availability and repair time is preserved. This identified dramatic shortfalls in availability for two of the FCC-ee operational modes [51].

3. **FCC-hh**: Like the LHC, the FCC-hh uses superconducting dipole magnets and RF cavities. Since the FCC-hh requires more than three times the amount of these components, as first approximation three times the downtime can be expected. Here it is assumed that the magnets in the LHC and the FCC-hh behave the same and that the downtime scales proportional. The downtime of the entire FCC-hh injector chain has to be considered in addition. If the LHC is used as an injector, its downtime of 106 days between 2016 and 2018, would certainly not be acceptable.

Additional innovations are required to meet the availability targets of CERN's next-generation accelerators. One possibility is to increase the reliability of the subsystems by extending the technological limits of their hardware components [53] or through installing redundancies. For example, the Super Proton Synchrotron (SPS) has been equipped with 1280 Solid State Power Amplifiers (SSPAs), for powering its RF cavities. The system carries on running, even if individual SSPAs stop working, which improves the availability of the SPS's RF system [54].

Another possibility is to decrease the repair time of faulty components. This can be achieved through modular designs for quick replacement and by optimizing the response time after a fault occurs. CERN further investigates several solutions for robot maintenance that allow components to be replaced during operation without downtime [55].

Fault prediction seeks to avoid downtime entirely by replacing components before they fail. This solution is considered one of the most promising to improve the availability in FCC-ee RF cavities [51]. In particular data-driven techniques are rapidly improving as sensors, data acquisition and storage becomes cheaper and more prevalent. The FCC-hh design report proposes data-driven fault prediction to overcome the availability challenge in the superconducting dipole magnets [5, p. 874]. Also in the CLIC implementation plan, data-driven fault prediction is proposed to reduce the downtime caused by RF cavity breakdowns [3, p. 35]. The next chapter discusses data-driven fault prediction in detail.

Chapter 3

Data-Driven Fault Prediction

Fault prediction can generally be achieved with physical models or data-driven models. Physical models aim to model the underlying physical processes of a system and its faults. If these physical processes are known, the system's operating behavior can be simulated and compared with actual behavior. In case there is evidence of a fault process, the exact cause of the fault can be determined, and the impact of corrective actions can be simulated. The CERN project Simulation of Transient Effects in Accelerator Magnets (STEAM) aims to generate such physical models based on the known physical processes of the superconducting magnets and their circuits [16]. When the physical processes are not known, it is possible to learn from past system behavior with data-driven models. Such methods are used in this thesis.

This chapter provides the mathematical notation of fault prediction data in Section 3.1, followed by an overview of data-driven models in Section 3.2. The latter section describes the categories of data-driven modeling and reviews models applied in the field of CLIC RF cavities and LHC main dipole magnets. For each category, a model relevant to this thesis is explained in detail with an example. The example demonstrates the prediction capabilities, the required system knowledge, and the interpretability of each model. Section 3.3 summarizes these characteristics. The foundation provided in this chapter should enable the reader to comprehend the terminology, requirements, and limitations of the algorithms used in the Chapters 4, 5, and 6.

3.1 Mathematical Notation

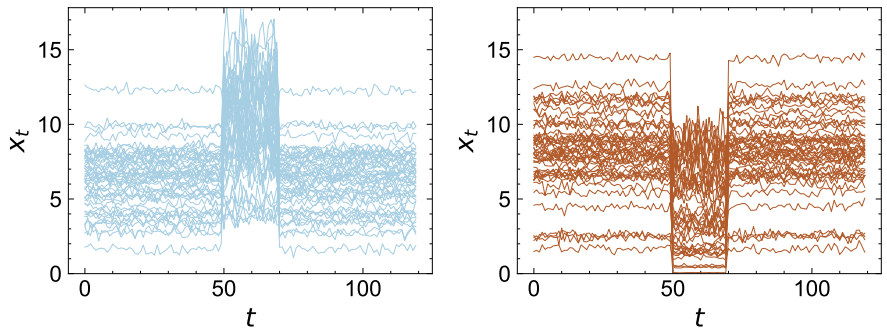
Definition 3.1.1 (Time-Series Signal)**.** A time-series signal $\mathbf{x} = \{x_1, ..., x_T\}$ is an ordered set of T real-valued data points $x_t \in \mathbb{R}$, where $t = 1, ..., T$. These time-series signals $\mathbf{x}$ are vectors, represented as bold lower case letters. Data point x_t are scalars denoted as non-bold lower case letters.

Definition 3.1.2 (Event)**.** Multiple time-series signals from different sensors, recorded at the same time, are referred to as an event $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$. Such an event consists of M time-series signals, $\mathbf{x}_m$, where $m = 1, ..., M$. The resulting matrix is represented with a bold upper-case letter $\mathbf{X}$.

Definition 3.1.3 (Features)**.** A feature is a characteristic property of an event. It is denoted with a "*", as a scalar x^*. A set of F features yields a vector $\mathbf{x}^* = \{x_1, ..., x_F\}$, where $f = 1, ..., F$. The process of calculating a feature is known as *representation learning*.

Definition 3.1.4 (Dataset)**.** A dataset defines a set of N events recorded at different points in time. If the system's fault behavior is known, a label y_n can be assigned to the event $\mathbf{X}_n$ or to its features $\mathbf{x}_n^*$, for $n = 1, ..., N$. For fault prediction, there are two classes $y_n \in [-1, 1]$: Faulty events are labeled with $y_n = -1$, while healthy events are labeled as $y_n = 1$.

An artificially created dataset is shown in Fig. 3.1. This dataset is intended to show the typical structure of data for fault prediction and helps to explain the algorithms in the next section. The values of the data points are exemplary for measurements in the CLIC RF cavities or LHC main dipole magnets and could correspond, to the field-emitted current or the magnet voltage. The dataset consists of $N = 50$ healthy events (Fig. 3.1a) and $N = 50$ faulty events (Fig. 3.1b). In each event, there is $M = 1$ time-series signal, with $T = 120$ data points. The next section will discuss how to distinguish these events in more detail.



(a) $y_n = 1$ (b) $y_n = -1$.

Figure 3.1: Artificially created dataset to explain data-driven modeling. Blue time-series signals on the left show healthy events, while brown time-series signals on the right show events indicating faults.

3.2 Data-driven Models

Data-driven fault prediction aims to predict the label of an event, defined as $\hat{y}_n$. The prediction should be generally applicable, even for events that are not included in the dataset. The generalization allows to automatically analyze large amounts of data for faults. The prediction is derived in two steps:

1. Representation Learning: In the context of fault prediction, features represent characteristic properties of a faulty event. When sufficient expert knowledge of the fault indicators are available, the features $\mathbf{x}_n^*$ can be directly calculated from an event $\mathbf{X}_n$. If this is not the case, the model has to *learn* the feature from the dataset.

2. Inference and Decision: The process of distinguish a healthy event from a faulty event consists of two steps. The first step is referred to as inference, where the probability $p(\hat{y}_n = -1|\mathbf{x}_n^*)$ of an event indicating a fault is derived. In the second decision step, a threshold is set on this probability to define whether the event is predicted as healthy ($\hat{y}_n = 1$) or faulty ($\hat{y}_n = -1$). The choice of this *decision boundary* impacts the risk of an event being misclassified as faulty (at a low threshold) or healthy (at a high threshold). Based on the system knowledge available, this decision boundary is manually chosen or learned from existing data. If the labels y_n are known, the derivation is called *supervised learning*, otherwise *unsupervised learning*.

Machine learning is a collective term, generally used for data-driven methods in which parameters are optimized automatically. In this work, the methods are divided into six categories, shown in Fig. 3.2. The methods are sorted descending by the required system knowledge. Green boxes represent manual calculations which require system knowledge, and red boxes show automatically performed calculations which require more data and complex models. In the following subsections, the modeling methods in this figure are explained in more detail.

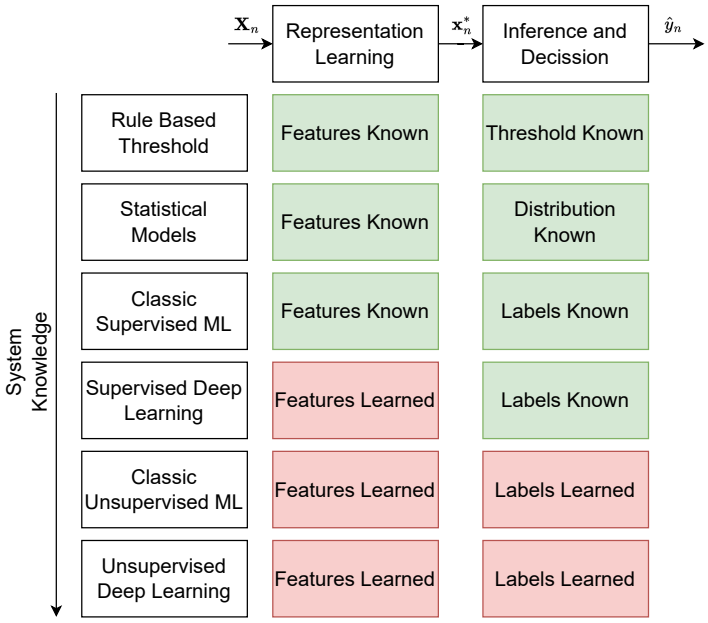


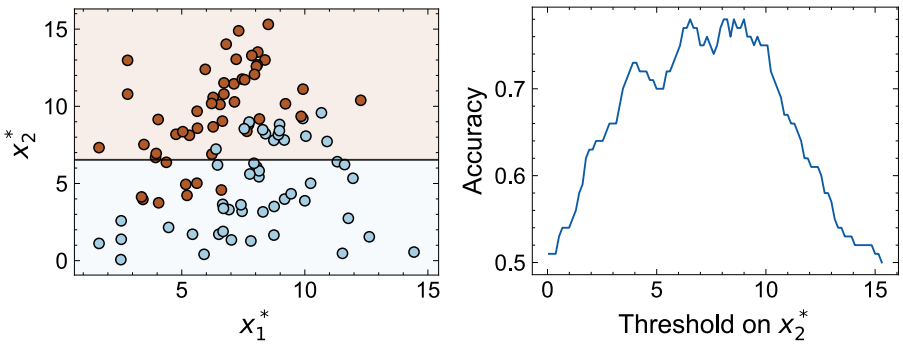
Figure 3.2: Overview of different data-driven methods for fault prediction. Green boxes represent manually performed calculations, while red boxes indicate automatically performed calculations.

3.2.1 Rule Based Thresholds

When the effect of a fault on the data is known, a simple threshold on a feature is sufficient for fault prediction. Both the feature and the threshold are chosen manually. In a CLIC test stand, such a threshold is applied to detect breakdowns based on the minimal value of the field emitted current [6]. Also in the LHC,

thresholds on the voltage measured in the superconducting magnets are used to detect quenches [7]. In these two applications, the thresholds are derived by experts, based on years of experience and a detailed understanding of the systems.

In addition to the system knowledge, data can support the experts to determine the threshold. In the artificially created dataset from Fig. 3.1, two features are indicating faults: The mean value x_2^* of the time-series signal in the range $t = [50, 70]$, and the mean value x_1^* of the remaining data points. For each time-series signal in Fig. 3.1, these features are calculated and shown in Fig. 3.3a. Blue dots represent healthy events, and the brown dots represent faulty events. The black threshold on x_2^* is selected, such that the ratio of correctly predicted events to all events, is maximal. This ratio is referred to as accuracy and plotted as a function of a threshold on x_2^* in Fig. 3.3b. Noticeably, the accuracy is highest with a threshold at $x_2^* = 6.5$ with 78%. The prediction with this threshold is visualized in Fig. 3.3a. The brown area marks faulty events. Brown dots in the blue area are not predicted correctly.



(a) (b)

Figure 3.3: Example of fault prediction with rule based thresholds. In (a) the brown area marks events indicating faults. Brown dots in the blue area are not recognized as faults. In (b) the threshold is selected, such that the accuracy to identify the event label is maximal.

This method is easy to interpret: The higher x_2^* the more likely the event is faulty. The closer x_2^* is to the threshold, the higher the uncertainty of the prediction. The method, can however only be applied if the exact features, indicating a fault are known. In addition, the prediction accuracy can be further improved with more complex thresholds, as the following subsection will show.

3.2.2 Statistical Models

Statistical models infer the probability $p(y_n = -1|\mathbf{x}_n^*)$ based on the distribution of features and select a threshold accordingly. This is how fault indicators in the field emitted current of CLIC RF cavities are identified [14]. In latest research, such thresholds are used to evaluate the LHC particle losses during a beam dump [56].

In more detail, the *prior knowledge* and *class-conditional densities* are required to infer $p(y_n = -1|\mathbf{x}_n^*)$. The *prior knowledge* refers to the probability of a healthy event $p(y_n = 1)$ and a faulty event $p(y_n = -1)$. This probability

can be derived from the data by taking the ratio of healthy or faulty events to all events. The *class-conditional densities* correspond to the distribution of the features for healthy and faulty events denoted as $p(\mathbf{x}_n^*|y_n = 1)$ and $p(\mathbf{x}_n^*|y_n = -1)$, respectively. These densities can be calculated with maximum likelihood estimation [57]. The desired probability $p(y_n = -1|\mathbf{x}_n^*)$ is referred to as posterior probability and can be derived with the Bayes theorem [58]:

$$p(y_n = -1|\mathbf{x}_n^*) = \frac{p(\mathbf{x}_n^*|y_n = -1)p(y_n = -1)}{p(\mathbf{x}_n^*)}, \tag{3.1}$$

where the denominator is calculated by $p(\mathbf{x}_n^*) = p(\mathbf{x}_n^*|y_n = 1)p(y_n = 1) + p(\mathbf{x}_n^*|y_n = -1)p(y_n = -1)$.

In the following, $p(y_n = -1|\mathbf{x}_n^*)$ is inferred for the artificially created dataset in Fig. 3.1. Both $p(y_n = 1)$ and $p(y_n = -1)$ are 0.5, as the number of healthy events is equal to the number of faulty events. The *class-conditional densities* are calculated with maximum likelihood estimation [57], assuming Gaussian distributions. The resulting standard deviations of these distributions are marked by the dashed ellipses of Fig. 3.4. Thus, all terms in Eq. (3.1) are available and $p(y_n = -1|\mathbf{x}_n^*)$ is inferred. The black continuous line in Fig. 3.4 shows the decision boundary, for $p(y_n = -1|\mathbf{x}_n^*) = 0.5$. With this decision boundary, 87% of all events are correctly classified.

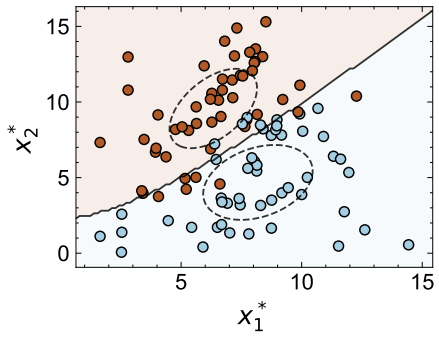


Figure 3.4: Example of fault prediction with a statistical model. The brown area marks events in which the probability of a fault is higher than no fault, while the blue area denotes a lower probability. The dashed ellipses represent the standard deviations of data with and without faults.

The advantage of this approach is that the probability $p(y_n = -1|\mathbf{x}_n^*)$ is determined and the accuracy is 9% higher compared to the rule based threshold. The probability helps system experts evaluate the confidence of the prediction. The *class-conditional densities* further show the features indicating a fault which makes the method interpretable. Estimating these densities for large datasets is computationally intensive. For small datasets, outliers in the data and changes in the system can significantly influence the results.

3.2.3 Classic Supervised Machine Learning

In classic supervised ML methods, the process of optimizing the decision boundary from section 3.2.1 is further extended. The methods vary from optimizing simple linear to complex non-linear decision boundaries [58].

Important algorithms used in this work are k-nearest neighbors [59], random forest [60] or Support Vector Machine (SVM) [61]. These algorithms are also used to detect faulty beam position monitors in the LHC [62] and at the Jefferson Laboratory [8] to detect faults in RF cavities. Specifically, in k-nearest neighbors, an event is classified by the majority class of its closest neighbors, typically measured using Euclidean distance. Random forests are ensembles of multiple decision trees, each of which applies multiple rule-based thresholds (see Section 3.2.1).

SVMs are used in both Chapters 5 and 6 of this thesis, and are therefore explained in more detail below. In an SVM the decision boundary is defined as:

$$f(\mathbf{x}_n^*) = \boldsymbol{w}^{\mathrm{T}}\phi(\mathbf{x}_n^*) + b, \tag{3.2}$$

where $\boldsymbol{w}$ are the weight parameters, b is the bias parameter and $\phi(\cdot)$ is a fixed feature space transformation. The feature space transformation allows modeling a non-linear decision boundary and corresponds to $\phi(\mathbf{x}_n^*) = \mathbf{x}_n^*$ for the linear case. The most common non-linear feature space transformation is the radial basis function. It is measured between two events $\mathbf{x}_n^*$ and $\mathbf{x}_n^{*'}$:

$$\phi(\mathbf{x}_n^*)^{\mathrm{T}}\phi(\mathbf{x}_n^{*'}) = \exp\left(-\frac{||\mathbf{x}_n^* - \mathbf{x}_n^{*'}||^2}{2\sigma^2}\right). \tag{3.3}$$

The parameter σ is referred to as hyperparameter, as it has to be manually chosen and is not automatically optimized by the algorithm. Support vectors, which give the method its name, are used for scaling $\boldsymbol{w}$ and b. They refer to events with the smallest perpendicular distance, i.e. the margin, to the decision boundary.

The optimization of $\boldsymbol{w}$ and b is performed by solving:

$$\begin{aligned} \arg\min_{\boldsymbol{w}} \quad & (\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{n=1}^{N}\xi_n) \\ \text{subject to: } & y_n(\boldsymbol{w}^{\mathrm{T}}\mathbf{x}_n^* + b) \geq 1 - \xi_n, \quad n = 1, ..., N \\ & \xi_n \geq 0, \end{aligned} \tag{3.4}$$

where ξ_n is referred to as slack variable, necessary to enable optimization, even if the dataset labels cannot be perfectly separated into $y_n = 1$ and $y_n = -1$ events. The hyperparameter C determines the importance of wrong predictions against the distance to the decision boundary. In general, a higher C leads to increased non-linearity of the decision boundary.

In Fig. 3.5 this method is used to predict faults in the artificially created dataset. Fig. 3.5a shows the prediction with a linear feature space transformation $\phi(\mathbf{x}_n^*) = \mathbf{x}_n^*$ and $C = 1$. This leads to a linear decision boundary, visualized by the black continuous line, predicting 89% of events correctly. The dashed lines represent the margin to the decision boundary defined by the support vectors. Fig. 3.5b shows the prediction of faults with an SVM using a radial basis function as a feature space transformation with $\sigma = 2$ and $C = 1$. The black continuous line indicates the non-linear decision boundary, and the dashed lines the margin. An accuracy of 92% is achieved.

This example shows that SVMs offer an opportunity to further improve

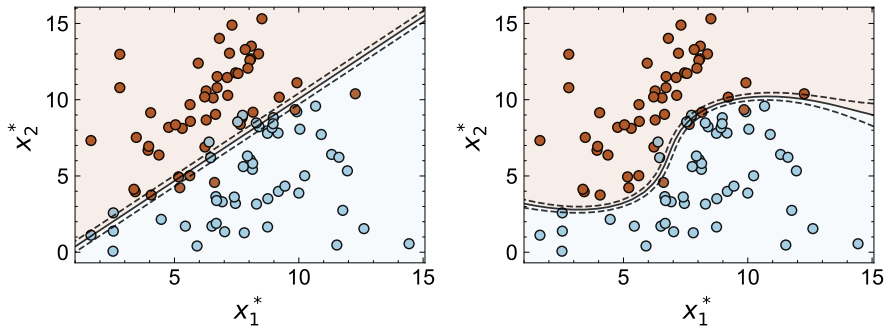


Figure 3.5: Example of fault prediction with SVM. In (a) a linear feature space transformation is used, in (b) a radial basis function. The black continuous line represents the decision boundary, and the black dashed lines show the margin.

the accuracy of fault prediction, compared to the previously described models. For the linear decision boundary, the weights $\boldsymbol{w}$ indicate which features are important for prediction. Important features can be validated by experts to ensure that the prediction is not based on bias. For the non-linear decision boundary in Fig. 3.5b it is not directly possible to understand which features are important for the prediction. Thus, the indicators on which the prediction is based cannot be validated without using additional interpretation methods.

3.2.4 Supervised Deep Learning

Deep Learning (DL) is a subfield of ML, which simulates the behavior of neurons in brain cells [63]. These neurons are sequentially chained together, resulting in a deep structure, which give the method its name. The deep structure allows learning data representation when the features indicating a fault are unknown. In the LHC quench heaters, such models have been investigated to detect faults [15]. At SLAC National Laboratory DL is used to predict beam properties [64].

A neuron represents a linear decision boundary, similar to Eq. (3.2). Instead of a feature space transformation $\phi(\cdot)$, however, an activation function $\sigma(\cdot)$ is used to create nonlinearity. The output of a neuron is therefore derived by:

$$f(\mathbf{X}_n) = \sigma\left(\mathbf{W}^\intercal \mathbf{X}_n + \mathbf{b}\right), \tag{3.5}$$

where $\mathbf{W}$ is the weight parameter, $\mathbf{b}$ is the bias parameter. The input of $\sigma(\cdot)$ is referred to as the activation a of the neuron. Typical activation functions include the sigmoid function given by $\sigma(a) = \frac{1}{1+e^{-a}}$ and the Rectified Linear Unit (ReLU) expressed as $h(a) = \max(0, a)$. The selection of an activation function is influenced by various factors, such as the convergence speed [63].

The standard deep learning model is the Neural Network (NN), in which neurons are grouped together in layers. In each layer, the output of the previous layer is used as an input. In a L-layered NN, the output is therefore given as

$$f(\mathbf{X}_n) = f_L(f_{L-1}(...f_1(\mathbf{X}_n)...)). \tag{3.6}$$

Note that in Eq. (3.6) the time-series signals $\mathbf{X}_n$ are used as an input. The output of each layer, represent features $\mathbf{x}_n^*$. In each layer, their calculation becomes more non-linear, and the distinction between healthy and faulty becomes easier. If a sigmoid activation is used in the last layer with one neuron, the output finally corresponds to $p(y_n = -1|\mathbf{X}_n)$, and faulty events can be determined with a simple threshold. The weights are then optimized to minimize the binary cross-entropy

$$L(y_n, \hat{y}_n) = y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n), \tag{3.7}$$

where $y_n = -1$ is substituted with $y_n = 0$

To show an example, faults in the artificially created dataset are predicted with a five-layer NN. The used network consists of three layers with 50 neurons, followed by one layer with two neurons and one neuron in the last layer. All layers use a ReLU activation function, except for the last layer, where a sigmoid activation function is used. With this model, an accuracy of 100% is achieved. Fig. 3.6 shows an example of the output x_1^* and x_2^* of the two neurons of the fourth layer. These features are transformed in a succession of non-linear layers, which explains the difference to features in the plots before. One can see that the healthy and the faulty events can be separated by the black decision boundary, set at $p(y_n = -1|\mathbf{X}_n) = 0.5$. The upper dashed line represents, $p(y_n = -1|\mathbf{X}_n) > 0.9$ while the lower dashed line shows $p(y_n = -1|\mathbf{X}_n) > 0.1$. It can be seen that most of the events are not between the two dashed lines, and are therefore predicted with high confidence.

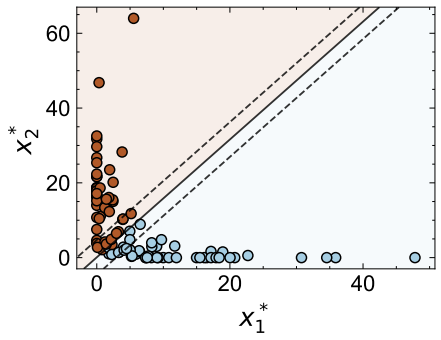


Figure 3.6: Example of fault prediction with an NN, where x_1^* and x_2^* are generated automatically. Events in the brown area are predicted as faults. The decision boundary indicates where the probability of a fault is higher than no fault. Events below the dashed line in the brown area are predicted as faults with 90% accuracy. Events above the dashed line in the blue area are predicted as no faults with 90% accuracy.

The achieved accuracy is 22% higher than the rule based threshold, 13% higher than the statistical model, and at least 8% higher than the SVM based fault prediction. This demonstrates the good fault prediction capabilities of deep learning. However, the prediction cannot be interpreted. In the example above, it is not clear which characteristic in the time-series signal x_1^* and x_2^* represent. Therefore, the indicators on the basis of which the prediction is obtained cannot be identified and validated by experts.

3.2.5 Classic Unsupervised Machine Learning

When no labels are available, the decision boundary is derived based on distinct features in the time-series signals. The representation learning task and the inference and decision task are then usually performed with separate methods, described in this subsection.

Representation Learning

Depending on the application, different unsupervised representation learning methods are used. In the LHC, Principal Component Analysis (PCA) is used to learn characteristic beam properties [65]. PCA successively transforms the data into orthogonal coordinate systems while aiming to preserve its variance [58]. At Daresbury Laboratory t-distributed Stochastic Neighbour Embedding (t-SNE) is used [66] to derive properties of time-series signals measured at the RF cavities. t-SNE aims to reconstruct the distribution of events in a lower dimensional space, by iteratively comparing the events in pairs [67]. In Chapter 6 Non-negative Matrix Factorization (NMF) is used for unsupervised representation learning and is therefore described in more detail below.

NMF decomposes the data as a linear combination of non-negative components [68]. The decomposed components are additive and are therefore easy to understand by humans. The input data of NMF is a two-dimensional matrix $\mathbf{V}$ with non-negative entries $v_{n,t}$ for $n = 1, .., N$ and $t = 1, .., T$. This matrix is decomposed into a $N \times K$ matrix $\mathbf{W}$ and a $K \times T$ matrix $\mathbf{H}$ such that:

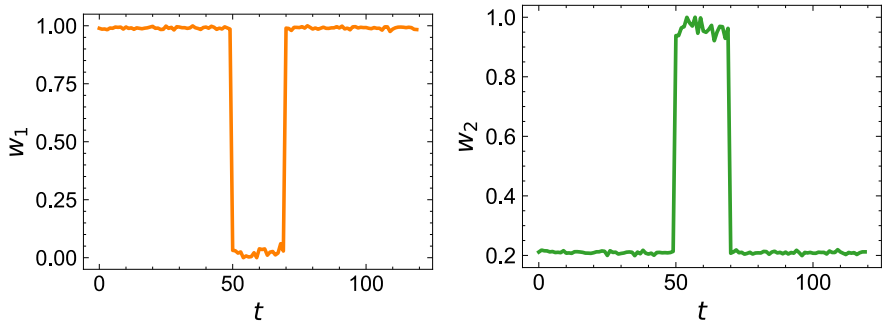
$$\mathbf{V} \approx \mathbf{WH}. \tag{3.8}$$

Here, $\mathbf{W}$ represents the components and $\mathbf{H}$ their weights. The parameter K defines the number of components. All elements $w_{i,k}$ and $h_{k,j}$ of the matrices $\mathbf{W}$ and $\mathbf{H}$ are constrained to be non-negative. To find the values of $\mathbf{W}$ and $\mathbf{H}$, both matrices are randomly initialized and optimized iteratively. In each iteration, the distance measure $d(\cdot)$ between the input $v_{n,t}$ and the reconstructed input $\hat{v}_{n,t} = \sum_k^K w_{n,k} h_{k,t}$ is minimized:

$$d(v_{n,t}, \hat{v}_{n,t}) = \|v_{n,t} - \hat{v}_{n,t}\|^2. \tag{3.9}$$

Further initialization methods and distance measures are described in detail in Chapter 6.

To apply this method, the artificially created dataset with dimensions $\mathbb{R}^{N \times M \times T}$ is transformed into a two-dimensional matrix $\mathbf{V} \in \mathbb{R}^{NM \times T}$. Two components $\mathbf{W}$ are extracted, visible in Fig. 3.7. They are summed to reconstruct the time-series signals in Fig. 3.1. Component two in Fig. 3.7a is used to model the period $t = [50, 70]$, and component one in Fig. 3.7b models the remaining data points. NMF can therefore reproduce the characteristics of the features in section 3.2.1, without labels or system knowledge. There, the mean value of the time-series signal in the range $t = [50, 70]$ was assigned to x_2^*, and the mean value of the remaining data points to x_1^*. The weights $\mathbf{H}$ of the components, are used to predict faults.



(a) (b)

Figure 3.7: The NMF components (a) and (b) extracted from the artificially created dataset. These components can be added together to reconstruct the events from the artificially created datasets.

Inference and Decision

In this step different methods are used dependent on the application. In the LHC, PCA features representing beam properties, are used to predict beam instabilities with isolation forest [65]. Isolation forest estimates the distribution of features and predicts events with features outside this distribution as faults [69]. At Daresbury Laboratory the features derived from t-SNE showing attributes of time-series signals from RF cavities are used to identify RF breakdowns with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [66]. DBSCAN identifies groups of events that are densely connected to each other, by measuring the distance between the features of all events [70]. In Chapter 4 a method called k-Means is applied, and therefore explained in more detail below.

k-Means [71] aims to assign N events to K groups defined as $\mathbf{S} = \{S_1, ..., S_K\}$. This is achieved, by randomly choosing a prototype $\boldsymbol{\mu}_k$ for each group S_k, where $k = 1, ..., K$. Each event $\mathbf{X}_n$ is assigned to the nearest $\boldsymbol{\mu}_k$ based on the Euclidean distance:

$$S_k = \{\mathbf{X}_n : ||\mathbf{X}_n - \boldsymbol{\mu}_k||^2 \leq ||\mathbf{X}_n - \boldsymbol{\mu}_j||^2 \; \forall j, 1 \leq j \leq k\} \tag{3.10}$$

The prototypes are then updated by the mean of all events in the group:

$$\boldsymbol{\mu}_k = \frac{1}{|S_k|} \sum_{\mathbf{X}_n \in S_k} \mathbf{X}_n, \tag{3.11}$$

where $|S_k|$ is the number of events in S_k. The steps from Eq. (3.10) and Eq. (3.11) are repeated until $\boldsymbol{\mu}_k$ remains constant.

In Fig. 3.8 the previously extracted NMF component weights $\mathbf{h_1}$ and $\mathbf{h_2}$ are visualized. By multiplying the NMF components in Fig. 3.7 with $\mathbf{h_1}$ and $\mathbf{h_2}$, each time-series signal in Fig. 3.1 can be reconstructed. One can see that the distribution of learned features $\mathbf{h_1}$ and $\mathbf{h_2}$ is similar to the distribution of manually chosen features $\mathbf{x_1^*}$ and $\mathbf{x_2^*}$ in Fig. 3.3a. This demonstrates the good performance of NMF.

The assignment of the events to a healthy and faulty group with k-Means is shown in Fig. 3.8. Events in the blue are assigned to the healthy group, events in the brown area to the faulty group. The large white circles show the k-Means prototype of each group, while the black line shows the decision boundary, where the distance from both prototypes is equal. 80% of all events are labeled correctly with this decision boundary.

Even if the accuracy is 20% lower than with the NN, this example shows that data-driven fault prediction is also possible if neither the features nor the labels of a dataset are available. The NMF representations are easy to understand and similar to the manually calculated features. Also, the k-Means prototypes are interpretable, as they show the mean of each group.

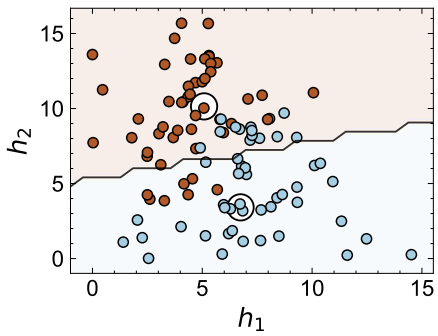


Figure 3.8: The weights $\mathbf{h_1}$ and $\mathbf{h_2}$ of the two NMF components for each event are grouped into healthy and faulty events with k-Means. Events in the blue area are considered to be healthy, events in the red area as faulty.

3.2.6 Unsupervised Deep Learning

Deep learning also allows for deriving non-linear representations, indicating a fault, even if no labels are available. The most frequently used model for this task is the Autoencoder (AE). Such an AE is used at Jefferson Laboratory to derive representations for fault prediction in CLIC cavities [66], [72]. At the Argonne National Laboratory, AEs are used to identify precursors of faults in magnets [73]. The method presented in Chapter 4, is also based on an AE.

An AE is an NN and therefore consists of sequentially interconnected neurons (Eq. (3.6)). As no labels are available the binary cross-entropy in Eq. (3.7) cannot be used to optimize the weights of the neurons. Instead, an AE aims to reconstruct the input $\mathbf{X}_n$ with the mean squared error:

$$L(\mathbf{X}_n, \hat{\mathbf{X}}_n) = ||\mathbf{X}_n - \hat{\mathbf{X}}_n||^2 \tag{3.12}$$

The number of neurons in the output layer matches the size of the input $\mathbf{X}_n$. The outputs of an intermediate layer, which typically has fewer neurons than the output layer, are then used as representations $\mathbf{x}_n^*$. The layers up to this intermediate layer are referred to as *encoder*, while the subsequent layers are the *decoder*.

An AE with three layers and 2, 50, and 120 neurons, is used to calculate the representations of the artificially created dataset. No activation function is used in any layer, leading to a linear function $f(\cdot)$ of the AE. The output x_1^* and x_2^*

of the first layer is shown in Fig. 3.9, where each dot shows an event colored in blue (healthy) and brown (faulty). The events are assigned to a group of healthy (brown area) and a group of faulty events (brown area) with k-Means. 6 brown dots are in the blue area and 12 brown dots are in the blue area, which means that 82% of events are correctly predicted.

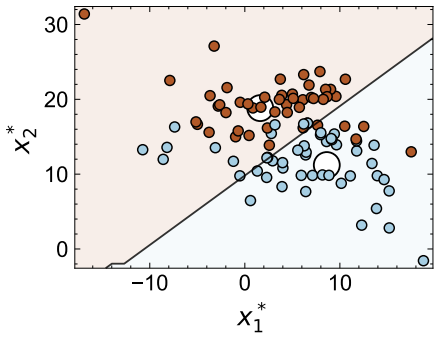
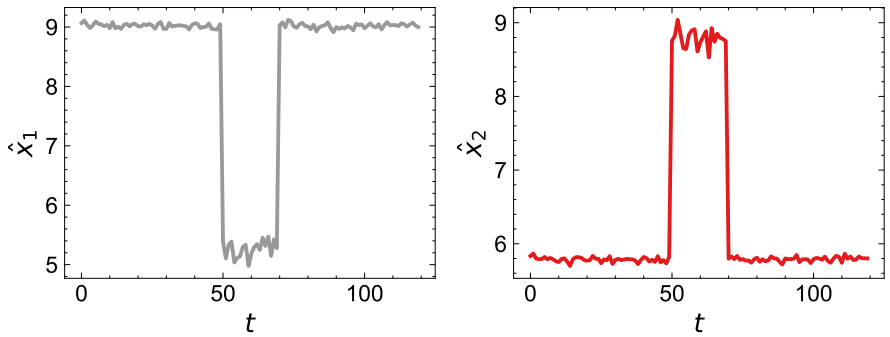


Figure 3.9: The output x_1^* and x_2^* of the first AE layer grouped into healthy and faulty events with k-Means. Events in the blue area are predicted healthy, events in the brown area as faulty.

Similar to NMF, the decoder of the AE allows reconstructing the time-series signals. As shown in Fig. 3.10, the two extracted k-Means prototypes can be reconstructed. These prototypes are representative of healthy events (Fig. 3.10a) and faulty events (Fig. 3.10b). This shows the feature indicating a fault: At $t = [50, 70]$ healthy events have a downward spike, while faulty events have an upward spike. Potentially also non-linear representations can be derived with non-linear activation functions, resulting in an increased flexibility of this method compared to NMF.



(a) (b)

Figure 3.10: The two k-Means prototypes in (a) and (b) reconstructed with the AE. The first component is representative of a healthy, while the second is representative of a faulty signal.

Model	Accuracy	Interpretable Results
Rule Based Thresholds	78%	✓
Statistical Model	87%	✓
SVM - linear	89%	✓
SVM - non-linear	92%	✗
NN	100%	✗
NMF/k-Means	80%	✓
AE /k-Means	82%	✓

Table 3.1: Summary of the characteristics of data-driven models applied on the artificially created dataset.

3.3 Summary of Data-driven Models

Table 3.1 gives a summary of all models applied to the artificially created dataset in the previous section. For each model, it is stated whether the result is interpretable. The accuracy is highest for complex supervised models: the non-linear SVM, and the NN. However, both methods are not interpretable, so the indicator on the basis of which the prediction is made cannot be determined. Therefore, a bias in the prediction cannot be excluded, and the reliability of the method cannot be assured. The next chapter will show how these models can still be interpreted with additional explanation methods.

Chapter 4

Interpreting Fault Predictions

As shown in the last chapter, supervised ML methods can predict faults accurately, but often lack interpretability. In fault prediction interpretability is beneficial to allow system experts to validate the results in the common case of few faults in the data. This chapter describes how to interpret supervised ML methods with Explainable Artificial Intelligence (XAI).

Existing XAI methods are developed to explain image or text predictions to non-experts [12], [74]–[76]. Those methods can also be used to explain fault predictions to system experts, as shown in Section 4.1. However, it is unclear which method is best suited for this application. Hence, Section 4.2 aims to answer RQ1: "How can system experts best interpret machine learning models to obtain reliable fault predictions?". This section summarizes the result from Paper 5: A novel XAI method for interpreting ML-based fault prediction tested on system experts.

4.1 Review of XAI Methods

Fig. 4.1 shows three different types of explanations, based on an event of the artificially created dataset in Fig. 3.1. Here, the goal is to explain why the event in Fig. 4.1a is predicted as faulty. In Fig. 4.1b, the negative spike is highlighted in pink, to show the area relevant for the prediction. This interpretation is referred to as *relevance-based* explanation. In contrast, the *concept-based* explanations show an event similar to the input that contain the characteristic concept of a faulty event: the negative spike. This event is chosen from the dataset and referred to as *example* explanations (Fig. 4.1c), or artificially generated and referred to as *prototype* explanation (Fig. 4.1d).

The above example explains the prediction of one specific event. This explanation is not suited for a different event, and is therefore referred to as *local* explanation. Explanations that are valid across all model predictions, are referred to as *global* explanations. Both types are discussed in more detail below.

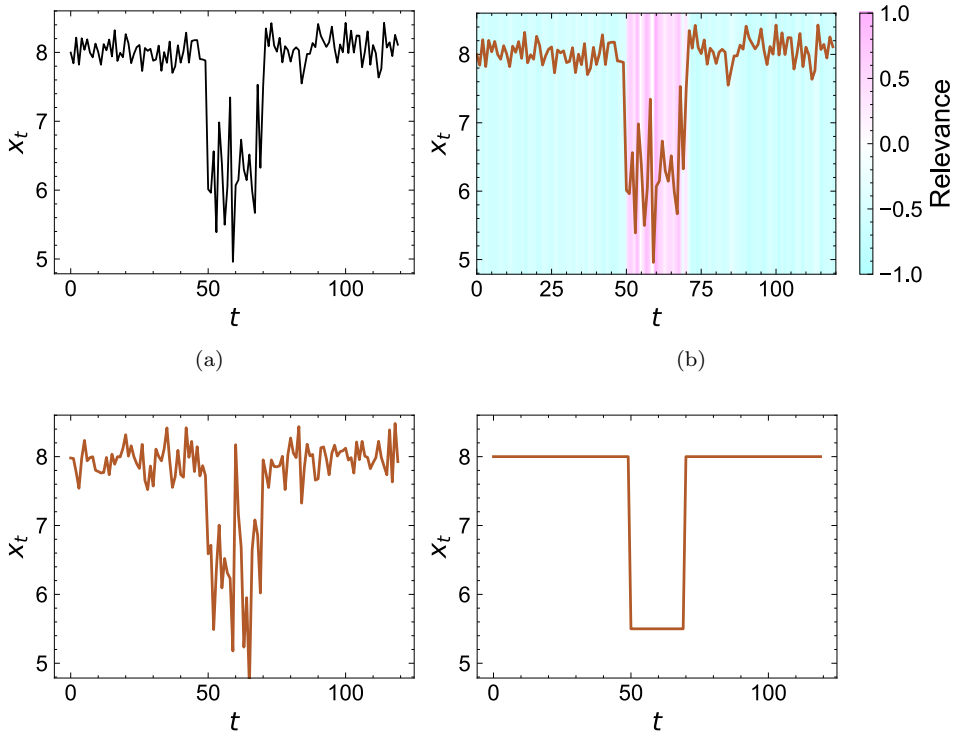


Figure 4.1: Different explanations to interpret a fault prediction from the artificially created dataset in Fig. 3.1. The goal is to explain why the raw data sample in (a) is predicted as faulty. In (b) relevance-based explanations highlight important parts, in (c) example explanations show similar events that are also faulty, and in (d) prototype explanations show an artificial event capturing the most important information.

Relevance-based Explanations

In a linear SVM (see Section 3.2.3), the weights $\boldsymbol{w}$ of the linear decision boundary $g(\mathbf{x}_n^*) = \boldsymbol{w}^{\mathrm{T}}\mathbf{x}_n^* + b$ indicate which features are important for prediction. A non-linear decision boundary $f(\mathbf{x}_n^*) = \boldsymbol{w}^{\mathrm{T}}\phi(\mathbf{x}_n^*) + b$ cannot be interpreted directly like this. Instead, the commonly used relevance-based methods LIME [75], Layer-Wise Relevance Propagation [77], and DeepLIFT [78] derive a local linear model $g(\mathbf{x}_n^*)$ behaves similar to $f(\mathbf{x}_n^*)$ for one event $\mathbf{x}_n^*$. For this event, the importance of the input for the prediction of the non-linear model is reflected in the weights of the local linear model.

Other commonly used relevance-based methods are saliency maps [79] or Grad-CAM++ [80]. These methods examine how the output behaves in response to a change in the input. If a change in a data point of the input influences the output, then it is relevant.

All these methods involve stochastic processes that can lead to varying results. With cooperative game theory [81] the above methods can be improved to obtain unique relevance values, referred to as SHapley Additive exPlanations (SHAP) values [76].

All these methods create local explanations, only valid for one specific event. To derive global methods, the relevant data points of several events are statistically evaluated [82], [83].

Concept-based Example Explanations

This method provides events that are similar to the event to explain, and is therefore also referred to as Explanation-By-Example (EBE). For fault prediction, EBE shows an example event where a fault already occurred. To find the example, the activations of the last layer in an NN are considered [12], [74], [84]. Events with similar activations share common properties that are important for the prediction (see Fig. 3.6). To explain a faulty event, another event with similar activations is therefore used as a local example. This similarity is frequently measured with the Euclidean distance [74], [84] or the cosine similarity [12]. Global explanations are derived, by grouping the activations with k-Means into a healthy and a faulty group (see Section 3.2.5). Events with similar activations to the prototype of the faulty group, are global examples for faulty events. The same principle is applied to explain healthy events.

Concept-based Prototype Explanations

Concept-based prototype methods aim to generate an artificial event, similar to the event to explain. For fault prediction, this artificial event aims to distill the main properties of a fault. In Fig. 4.1d this property is the negative spike.

Generating an artificial event is possible with an AE, discussed in Section 3.2.6. Concept-based prototype methods therefore frequently rely on AEs [85], [86]. This AE is part of the NN and aims to reconstruct the activations of the last layer. The reconstructions are used as local explanations. If the activations are grouped into healthy and faulty, as in concept-based example method, the prototypes can be reconstructed for each class. The reconstructed prototypes are global explanations. Since the AE is part of the network, the approach is referred to as Model-Specific Prototype (MSP) method. This has the disadvantage, that the method cannot be used if the model structure is unknown.

For both concept-based methods, a combination with relevance-based explanations is possible [83]. For this purpose, important parts are shown in the example [87]–[89] or prototype [90]. As these important parts frequently represent shapes in the time-series signal, they are referred to as shaplets. In Chapter 5 these shaplets are used to explain the fault predictions of CLIC RF cavities.

A recent study shows that users prefer concept-based explanations for time-series predictions [12]. This makes these explanations useful for fault prediction, where time-series data are commonly used. However, there are no studies testing concept-based explanations for system experts. In the next section, a proposal to address this limitation is discussed.

4.2 Research Contribution 1

In the context of this research contribution, a novel XAI model is proposed for explaining any fault prediction model with concept-based prototypes. This XAI model is compared to two other concept-based XAI methods in a human-user study. The results are presented in detail in Paper 5 and summarized below.

Model

The proposed XAI model is based on an AE, similar to other methods. As it can explain unknown black box models, it is referred to as Model-Agnostic Prototype (MAP) method. Fig. 4.2 shows the structure of the MAP method. The input $\mathbf{X}_n$ is reconstructed with an encoder $g(\cdot)$ and a decoder $h(\cdot)$. The reconstructed input $\hat{\mathbf{X}}_n$ is then used as an input for the black box model $f(\cdot)$, which makes class predictions. To obtain explanations, the output of the encoder is assigned to one of K groups with k-Means. Subsequently, the decoder reconstructs the k-Means prototypes $\mathbf{c}_k$, where $k = 1, ..., K$. These reconstructed prototypes are used as concept-based explanations for the prediction of the black box model.

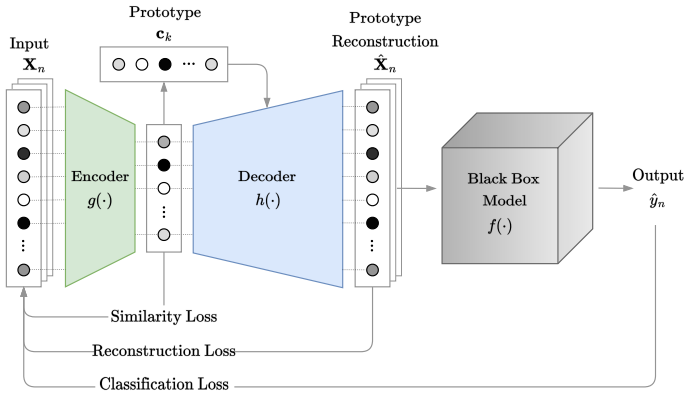


Figure 4.2: Model architecture of the MAP explainer. Given a trained black box model, an AE is fitted to reconstruct the input data and to recreate the output of the black box model. The prototypes are derived from the output of the encoder and are optimized to differ from each other.

The optimization of the AE weights is performed by minimizing three different losses. To reconstruct the input $\hat{\mathbf{X}}_n$ as accurately as possible, the mean-squared-error is used:

$$R(g, h, \mathbf{X}_n) = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{X}_n - h\left(g(\mathbf{X}_n)\right)\right)^2 . \tag{4.1}$$

The label $\hat{y} = f\left(h\left(g(\mathbf{X}_n)\right)\right)$ should reconstruct the output of the black box model $f(\mathbf{X}_n)$. The prediction accuracy is measured by the categorical-cross-entropy loss:

$$C(g, h, f, \mathbf{X}_n) = -\sum_{n=1}^{N} \left(\underset{\hat{y}}{\operatorname{argmax}}\, f(\mathbf{X}_n)\right) \log f\left(h\left(g(\mathbf{X}_n)\right)\right) . \tag{4.2}$$

A similarity loss ensures that the derived prototypes are differing from each other:

$$S(\mathbf{C}) = \frac{\sum_{i \neq j} \mathbf{c}_i^T \mathbf{c}_j}{K(K-1)}. \tag{4.3}$$

This similarity loss, is penalizing non-orthogonality between two different prototypes $\mathbf{c}_i, \mathbf{c}_j \in \mathbf{C}$, where $\mathbf{C} = \{\mathbf{c}_1, ..., \mathbf{c}_K\}$ [91]. Overall, the weighted sum of all three losses,

$$\mathcal{L}(g, h, \mathbf{X}_n) = R(g, h, \mathbf{X}_n) + \lambda_C C(g, h, f, \mathbf{X}_n) + \lambda_S S(\mathbf{C}), \tag{4.4}$$

is optimized. The weights λ_C and λ_S are hyperparameters. To evaluate the effectiveness of this method, the MAP model is compared to a concept-based EBE [12] method and a concept-based MSP method [86]. These two methods are chosen because they are commonly used for concept-based explanations (see Section 4.1). The architectures of all three methods are discussed in detail in Paper 5.

Data

Eleven dataset from the University of California Riverside (UCR) archive [92] are compared in a quantitative analysis. Two datasets are additionally evaluated with a user study, which will be summarized in this section. Specifically, those datasets are:

1. **ECG200:** The ECG200 [92], [93] dataset contains data of electrical activity measured during one heartbeat. Specifically, the last part of a heart beat is shown in the signal (see Fig. 4.3), starting after the peak point R. The characteristic properties of a normal heart beat (healthy event) compared to an ischemic heart beat (faulty event) are the high peak point R and the limited recovery time from its minimum S to T. The scaled reconstruction of the ground truth is shown in Fig. 4.3 together with the characteristic points R, S, T, and U.

2. **Artificial Dataset:** Furthermore, an artificial dataset is created, which aims to reconstruct time-series signals from machine sensors in a noisy environment. Four basic time-series shapes represent the ground truth signals shown in Fig. 4.4. Multiplicative and additive noise with an amplitude of 0 to 1.1, which is drawn from an uniform distribution, is added to these ground truth signals.

Results

Fig. 4.3 and Fig. 4.4 show explanations generated with the EBE, the MSP, and our MAP method for the ECG200 and the artificial dataset, respectively. For each method, one survey is created[1]. As part of the survey, two global explanations per class were shown to a group of participants. Afterwards, this group had to predict the label of 15 randomly drawn events from each dataset. The ratio of correctly predicted events is used as a performance measure.

[1] EBE: `https://forms.gle/J3EAnAqN99mpw6P39`
MPS: `https://forms.gle/rsRzHcXyurPi6LQA9`
MAP: `https://forms.gle/tSZRXbuZUraKW7cz8`

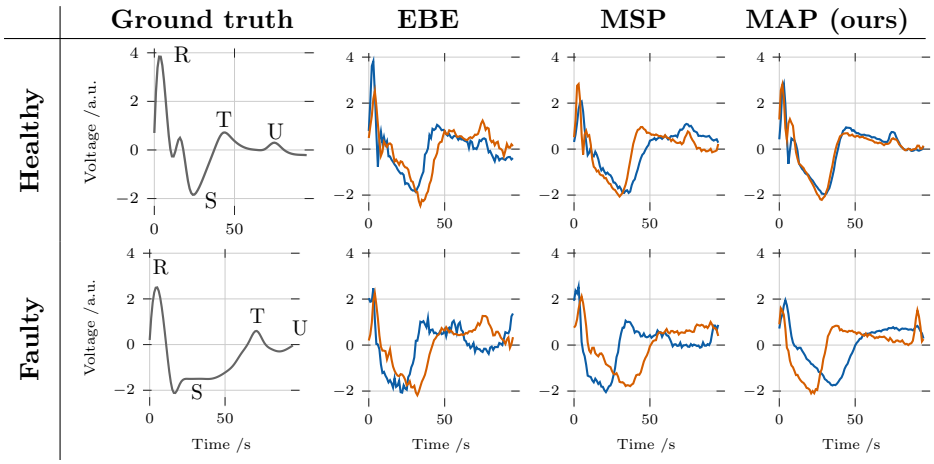


Figure 4.3: Ground truth and explanation of the ECG200 [93] dataset, showing the latter part of a heart beat, starting before the peak R. For both healthy and faulty events, two explanations are extracted with the EBE, MSP, and our MAP method.

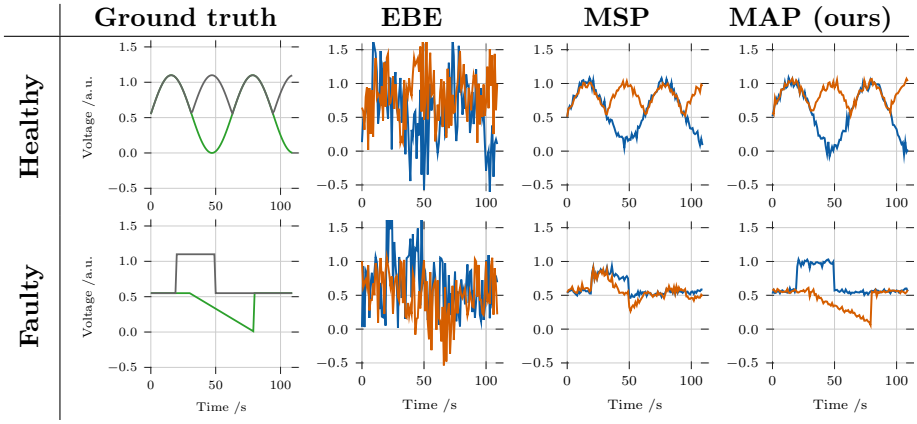


Figure 4.4: Explanations of artificially created dataset with two explanations for both healthy and faulty events, extracted with the EBE, MSP, and our MAP method. The ground truth signal shows the four shapes within the dataset, to which multiplicative and additive noise with an amplitude of 0 to 1.1, drawn from a uniform distribution, is added.

The surveys were distributed within our research community, collecting 2190 valid answers from 73 students and research staff from CERN and Graz University of Technology. The results differentiate between two categories of participants: At the beginning of the survey, participants who indicated prior ML knowledge are categorized as typical ML developers. Participants without prior ML knowledge were classified as ML users.

The study was evaluated based on the principle of binomial proportion [94]. Table 4.1 shows the results in the form $\hat{p} \pm z\sqrt{(\hat{p}(1-\hat{p}))/n}$. The first term $\hat{p}$ shows the proportion of success in the binomial trial, which indicates how many of the 30 events were on average correctly predicted by each participant. The latter term $z\sqrt{(\hat{p}(1-\hat{p}))/n}$, indicates the uncertainty of the result. z is the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution, where α is the target error rate. For

Participants	Method	ECG200 [%]	Artificial data [%]	Total [%]
Developer	EBE	65.6 ± 6.7	79.0 ± 5.7	72.3 ± 4.4
	MSP	64.4 ± 8.1	74.1 ± 7.4	69.3 ± 5.5
	MAP (ours)	**74.8 ± 5.2**	**84.8 ± 4.3**	**79.8 ± 3.4**
User	EBE	71.3 ± 7.3	**80.0 ± 6.4**	75.7 ± 4.9
	MSP	64.3 ± 6.5	67.6 ± 6.3	66.0 ± 4.5
	MAP (ours)	**77.8 ± 7.0**	78.5 ± 6.9	**78.1 ± 5.0**
All participants	EBE	68.1 ± 4.1	79.4 ± 3.1	73.8 ± 2.4
	MSP	64.3 ± 4.0	70.1 ± 3.2	67.2 ± 2.4
	MAP (ours)	**75.8 ± 3.6**	**82.7 ± 2.9**	**79.3 ± 2.2**

Table 4.1: Results of 73 participants predicting a total of 2190 events with explanations generated by EBE, MSP or our MAP method. The proportion of success is indicated alongside the 95% confidence level.

the chosen 95% confidence level, the target error rate is $\alpha = 0.05$.

Evaluating the answers of all participants from both datasets, our MAP method achieves a leading 79.3% success rate, compared to 73.8% of EBE and 67.2% of MSP. The MAP confidence intervals are not overlapping with those of the EBE or the MSP method, confirming the statistical significance of the results.

These results demonstrate that prototype explanations are preferred by the participants across the two datasets. Prototypes contain only the essential properties, which help the participants to focus on the relevant details, allowing an average success rate of 79.3% with MAP explanations. If prototypes are too different to the ground truth, as is the case with the chosen MSP method for both dataset, participants are misled and their success rate drops on average by 12.1%. In the artificially created dataset, users reached a 1.5% higher accuracy with EBE compared to our MAP method. This shows that when the prototype is too abstract compared to the real data, users of ML methods prefer examples over prototypes. Developers do not have this problem. Similar to other surveys [12], [76], this points out that explanations are perceived differently by the target groups. It is therefore important to take into account the target audience when using XAI methods.

Chapter 5

Case Study 1: CLIC Radio Frequency Cavities

Breakdowns are a key limitation for reaching a high accelerating gradient in CLIC RF cavities. They reduce the availability of the cavities, as the operation is interrupted after a breakdown, and degrade the linearly accelerated beam, rendering it lost on that pulse. Furthermore, they damage the cavity surface as shown in the microscopic image in Fig. 5.1. In this image, the remaining crater of molten copper after a breakdown is visible. A high number of craters can negatively impact a cavity's performance and can lead to the replacement of the cavity (see Section 7.5.3 of [95]). Replacing a cavity increases the downtime of the accelerator and is associated with high costs. A CLIC RF cavity costs approximately 75-95 kCHF [41].

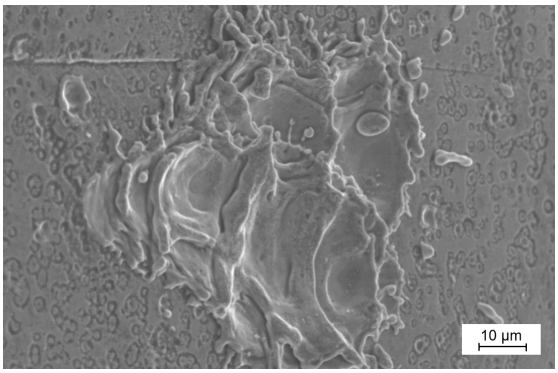


Figure 5.1: Crater of molten copper that forms after a breakdown [96].

Existing methods aim to monitor breakdowns, to detect and mitigate their negative effects. These methods are discussed in Chapter 5.1. Breakdown prediction could fully prevent the negative effects. Therefore, in Chapter 5.2 the question "Can data measured at CLIC RF cavities provide insights on breakdown prediction?" is answered by showing selected results from Papers 2, 4, and 3. This chapter summarizes the second research contribution of this thesis.

5.1 RF Cavity Breakdowns

RF cavity breakdowns emerge from surface deformations or contaminations that result in the local enhancement of the surface electric field. The increased electric field cause electron emissions, which can lead to a plasma that is referred to as breakdown. Existing methods for breakdown detection monitor the current emitted from RF cavities. Depending on how frequently breakdowns occur, the gradient in the cavity is then reduced accordingly [97].

In the CERN XBOX2 test stand for CLIC RF cavities, the same principle is applied. The electron emissions are measured upstream and downstream of the cavity with sensors known as Faraday Cups (FC). If emissions exceed a threshold, the operation is stopped for a few seconds.

Fig. 5.2 shows an example of the emissions measured in the downstream FC of the XBOX2 test stand during a RF pulse without (a) and with (b) a breakdown. The emissions are shown relative to the maximum resolution of the Analog to Digital Converter (ADC) in the FC. In the faulty RF pulse, emissions are clearly higher, and reach the ADC saturation after around 1.25 μs. At the XBOX2 test stand, a breakdown is detected if emissions reach 81.3% of the ADC saturation. With this threshold, the events in Fig. 5.2a and Fig. 5.2b can be clearly separated from each other.

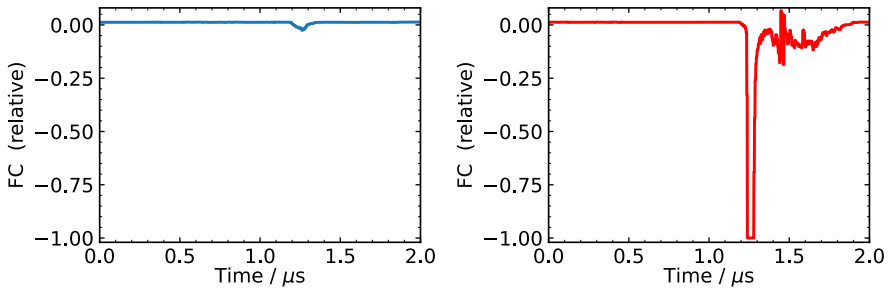


Figure 5.2: Two examples of a FC signal without (a) and with (b) a breakdown. Emissions are shown relative to the maximal resolution of the analog to digital converter in the FC.

Breakdowns usually occur in groups. The edges of a breakdown crater (see Fig. 5.1) represent surface deformations that increase the probability of another breakdown shortly thereafter [98]–[100]. To avoid that these *follow-up* breakdowns lead to even more craters and breakdowns [95, Fig. 7.31], the gradient is slowly increased after the *primary* breakdown in the XBOX2 test stand [95], [101]. Similarly, in the accelerators LINAC4 [102, slide 6] and CLARA [103, p. 61] the occurrence of multiple follow-up breakdowns is prevented through an automatic recovery procedure, that temporarily decreases the gradient in the cavity depending on the rate of breakdowns.

Breakdown prediction would allow preventing plasma formation by stopping operation beforehand, or by powering an additional spare cavity. The data measured at the XBOX2 test stand could provide additional insight into achieving this goal. This is further discussed in the next section.

5.2 Research Contribution 2

The analysis of the data measured at the XBOX2 is published in three papers. Paper 2 establishes the approach for ML-based breakdown prediction in RF cavities. Paper 4 applies this approach on two different XBOX2 datasets and evaluates the prediction performance of various ML methods. To further improve this prediction performance, Paper 3 discusses methods for artificially augmenting XBOX2 data. In this chapter, selected results of these publications are presented, summarizing the second research contribution.

Data

Every pulse at 50 Hz, the electron emissions in the cavity (see Fig. 5.2) and the power amplitude of the waves traveling to the cavity (see Fig. 3 in Paper 4) are recorded. These data form an event $\mathbf{X}_d$, and consist of time-series signals that are 2 μs long with up to 3200 data points. Although every pulse is monitored, not every event is stored due to the limited data storage of the experimental setup. Instead, recordings of one RF pulse are stored every minute. In addition, in the event of a breakdown, the recordings of the current and the two previous RF pulses are stored.

Prediction

The goal of this work is to predict, given an event $\mathbf{X}_d$, whether a breakdown will occur in the next pulse in 20 ms. For this purpose, the events that are stored every minute and the events one pulse before a breakdown are used. The former are labeled as healthy ($y_d = 1$) if no breakdown occurs in 20 ms, the latter are labeled faulty ($y_d = -1$) if a breakdown occurs in 20 ms.

Unlike the events in Fig. 5.2, the difference between healthy and faulty events is not evident, and they cannot be distinguished with a simple threshold. Therefore, the supervised ML and DL models from Chapter 3 are used. Their exact choice of hyperparameters is discussed in Paper 4.

The prediction performance of the models is evaluated with the Area under the Receiver operating characteristics curve (AR) [104]. This score evaluates the model capabilities to infer $p(\hat{y}_d|\mathbf{X}_d)$, independent of the threshold chosen in the decision step (see Fig. 3.2). For this purpose, the correctly predicted faulty events are calculated as a function of the incorrectly predicted faulty events for all possible thresholds. An AR score close to 100% indicates a clear discrimination between healthy and faulty events, while a score near 50% shows that the model selects the label randomly.

To assess the model performance, *leave-one-out-cross-validation* is used [105]. This means that the dataset is divided into five groups, where each group contains events with similar operational parameters. A model is then alternately trained on four groups and validated on the remaining group. Once each group has been validated, the average AR_μ and the standard deviation AR_σ of each validation is reported. After fine-tuning each model's hyperparameters, the model is trained on all five groups and evaluated on an additional test group, which was not used before. The prediction performance of the test group is reported as AR_t.

Tab. 5.1 shows the best results reported in Paper 4 for predicting primary and follow-up breakdowns 20 ms in advance. As primary breakdowns are

generally considered stochastic [106], the models' performance of $AR_\mu = 56.6\%$ is still slightly better than the expected 50 %. Nonetheless, the performance of $AR_\mu = 89.7\%$ for predicting follow-up breakdowns is much higher. Intuitively, this means that in 89.7% of all cases, the model correctly ranks a faulty event as more probable to be faulty than healthy. According to this performance, the prediction of breakdowns 20 ms in advance is possible for follow-up breakdowns, but challenging for primary breakdowns.

The question is on which indicator the follow-up breakdown prediction is based and if this indicator is reliable. Tab. 5.1 already provide insights to this question. The relatively high standard deviation of $AR_\sigma = 8.1\%$ shows that the performance varies in the validation groups. This implies that the fault indicator used by the model is not equally prominent in all groups. Nevertheless, the model manages to generalize well, as $AR_t = 91.1\%$ is higher than $AR_\mu = 89.7\%$. It shows that the fault indicator is also present in the test group.

Table 5.1: Best AR scores for predicting RF cavity breakdowns.

	AR_μ	AR_σ	AR_t
Primary Breakdowns	56.6%	8.3%	54.0%
Follow-up Breakdowns	89.7%	8.1%	91.1%

Interpretation

To gain further insights, the fault indicator itself is investigated with the XAI methods from Chapter 4. Specifically, the relevant time-series signal and its data points are evaluated with SHAP values [76]. Furthermore, two concept-based prototypes, representative of healthy and faulty events, are extracted with the method presented in Chapter 4. [1]

SHAP values show, that electron emissions in the cavity are more relevant to the prediction than the power amplitude of the waves traveling to the cavity (see Fig. 15 (a) of Paper 4). The reason why these emissions are important are shown in Fig. 5.3. In this figure, a prototype of a healthy signal is shown in blue, and a prototype of a faulty signal is shown in brown. For better comparability, both prototypes are standardized by subtracting them with the mean and dividing them by the standard deviation afterward. The data points of the prototypes which are relevant for predicting faulty events are highlighted in purple.

While the blue and the brown prototypes in Fig. 5.3 cannot be distinguished as easily as the signals in Fig. 5.2, they vary in several areas. The brown time-series signal is generally more noisy, has higher amplitude in the interval [1.25; 1.5], but lower amplitude in the interval [1.55; 1.9] compared to the blue signal. Only the first interval within [1.25; 1.5] is marked in purple, and therefore relevant for the prediction. In this interval, the brown signal shows fluctuations at 1.2 μs and at 1.4 μs, which are especially purple. These fluctuations appear to be the fault indicators for the prediction.

In a recent thesis, the identified interval [1.25; 1.5] was studied in detail [107]. As the RF pulse enters the individual subsections of the RF cavity, they emit electrons until the end of the RF pulse. These emissions form the main negative

[1] In Paper 4 concept-based examples are used, as the presented concept-based prototype method has been developed retrospectively.

spike in the interval [1.25; 1.5] independent of healthy and faulty events. The thesis also shows, that fluctuations in the dark current signal may be associated with an increased probability of follow-up breakdowns in the next pulse, i.e. faulty events. The same phenomenon is observed in Fig. 5.3. The brown signal, representative of faulty events, shows fluctuations at 1.2 μs and at 1.4 μs. These fault indicators therefore are compatible with previous observations. The prediction appears to rely on true fault indicators rather than a bias.

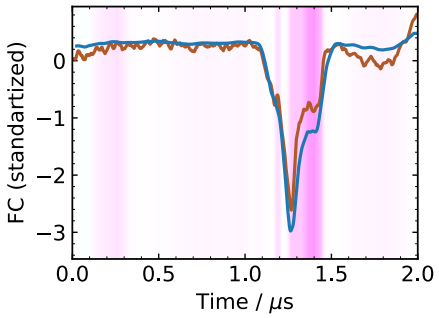


Figure 5.3: Interpretation follow-up breakdown prediction with XAI. The blue time-series signal represents the concept-based prototype explanation of a healthy event. The brown time-series signal represents a concept-based prototype explanation of a faulty event, in which a breakdown occurred in the next pulse 20 ms later. In addition, the purple area marks the region which is relevant to the model for the prediction.

Chapter 6

Case Study 2: LHC Main Dipole Magnets

Due to the high nominal energy of 1.1 GJ stored in each of the eight LHC superconducting dipole circuits, faults in superconducting components can have severe consequences [29]. These consequences can be seen in Fig. 6.1, which shows a hole burnt in the coil of a superconducting dipole magnet caused by an intermittent short circuit. Such a fault can lead to significant LHC downtime of up to three months, as the circuit has to be warmed up for repair and subsequently cooled down again in a sophisticated procedure. On top of the downtime, material costs account to around 1 MCHF for a LHC main dipole magnet [108].

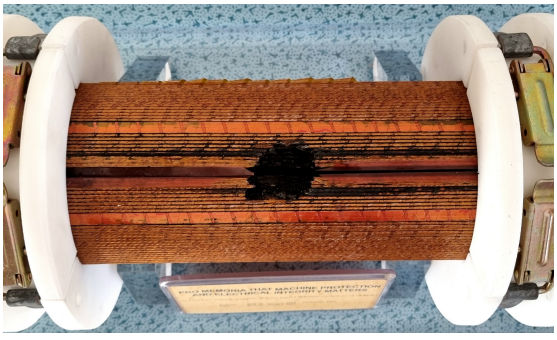


Figure 6.1: A hole burnt into the coil of a LHC superconducting dipole magnet after an intermittent short circuit show the potential negative consequences of a fault.

System experts therefore closely monitor the behavior of superconducting components, as described in Section 6.1. ML-based fault prediction could further help the experts to detect anomalies and automate this monitoring process. In this regard, Chapter 6.2 shows selected results from Paper 6 and 1 to answer the question "How can the data measured at the LHC main dipole magnets provide insights about their normal and abnormal behavior?". These results represent the third research contribution of this thesis.

6.1 Fault Protection in LHC Main Dipole Magnets

The resistive transition in a superconducting magnet, also called quench, is accompanied by local heating in the superconducting cables and high voltage transients in the magnet. The magnets and their protection systems are designed to withstand a quench, but in rare cases faults may still occur. To understand possible faults and how the magnets are protected against them, the LHC main dipole circuits and its quench protection systems are explained in this section.

Fig. 6.2 shows a schematic view of a main dipole circuit with its 154 magnets, each represented by a magnet inductance L_{M} [109]. For this analysis, the magnets are counted either along their physical position from P1 to P154 or clockwise along the electrical connection from E1 to E154. The Current Leads (CL) ensure the electrical connection between the cold superconducting part of the circuit and the warm, normal conducting parts of the circuit.

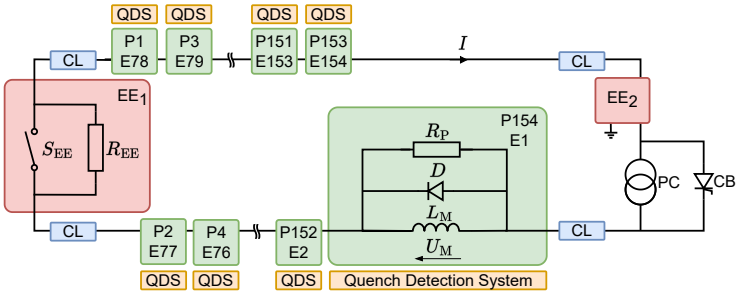


Figure 6.2: Schematic view of the main dipole circuit, including the Power Converter (PC), the Crowbar (CB) and the CL. The QDS triggers a FPA, which deactivates the PC and activates the energy extraction systems. Furthermore, it triggers the discharge of the QHs in the respective magnet, if a quench is detected. The two Energy Extraction Systems EE_1 and EE_2 consist of a Switch S_{EE}, and an Energy Extraction Resistance R_{EE}. The circuit is grounded at the center of the resistor R_{EE} in the EE_2 system. The magnet with inductance L_{M} and the by-pass Diode D with a Parallel Resistance R_{P} are in a liquid helium cryostat. Magnets are labeled by their Physical position (P) from the left to the right. The Electrical positions (E) are counted clockwise along the electrical connection starting from the PC. The numbering shown here is representing the circuits in sectors 12, 34, 56, and 78. In sectors 23, 45, 67, and 81 the electrical labels change, as the PC is on the left side of the circuit.

In case of a quench or other powering failures in the circuits, a system of protection elements is in place to safely dissipate the energy in the quenched magnets and extract the remaining energy of the circuit [110]. This process is referred to as a Fast Power Abort (FPA) event. The quench protection elements include a Quench Detection System (QDS), which detects the voltage increase due to a quench and triggers the appropriate protection actions [111], [112]. Upon the detection of a quench in a dipole magnet, the Power Converter (PC) is switched off and the current I by-passes the PC via the Crowbar (CB). Also, the Quench Heaters (QHs) of this magnet are activated. QHs are resistive strips attached to the outer surface of each magnet coil [113]. Upon activation, they heat up and cause the majority of the magnet coils to get normal-conducting in a few tens of milliseconds [15]. This ensures protection by distributing the magnet's stored energy uniformly over the quenched magnet windings [114]. The by-pass Diode D diverts current from the quenched magnet. This restricts

the quenching magnet to only absorb its stored magnetic energy, not the energy of the entire circuit. The Parallel Resistance R_{P} installed across each magnet, smoothens transient voltages during this process [115]. To avoid the circuit's energy to solely discharge in the diode of the quenched magnet, the switches S_{EE} in both Energy Extraction (EE) systems are sequentially activated [116]. They direct the circuit current towards the Resistances R_{EE}, which extracts the circuit's energy within around 300 s.

The voltages U_{M} measured over the 154 magnets during a single FPA event are shown in Fig. 6.3. These voltage transients contain information about the behavior of the electrical circuit and its components [117].

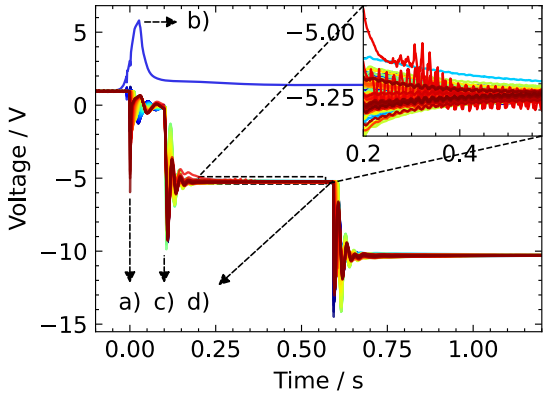


Figure 6.3: Voltages U_{M} across the 154 main dipole magnets of sector 78 following a quench in the magnet with the electrical position 141 on 31.03.2021 with its different phases: a) the FPA is triggered at 0 s, the QHs for magnet 141 are activated, and the PC is deactivated; b) after around 0.03 s the by-pass diode of the quenched magnet becomes conductive; c) the first Energy Extraction system EE_1 is activated about 0.1 s after the FPA trigger; d) the second Energy Extraction system EE_2 is activated approximately 0.5 s after the first one. The blue curve shows the voltage across the quenched magnet, while the remaining curves represent the voltages across the other 153 magnets of the circuit.

Fig. 6.4 shows schematically how the resistive strips of the QHs are mounted on the outside of a magnet coil. This figure shows the two High Field (HF) QHs and the two Low Field (LF) QHs for one of the two coils in a dipole magnet. The LF QHs remain electrically floating and serve as redundancy if the HF QHs fail. The QH circuit is powered by a capacitor bank which discharges it's energy of 2.86 kJ into two resistive QH strips connected in series within $\sim$ 300 ms. The voltage and the current measured during this discharge are shown in Fig. 6.5. Due to the current flow in the QH strips, they heat up and their resistance increases, resulting in a pseudo-exponential decay.[15]

To avoid faults in the LHC superconducting dipole magnets and their protection systems, they are extensively tested during Hardware Commissioning (HWC) and after each extended technical stop, before operation continues [118]. In addition, an annual maintenance period offers time to conduct measurements on hardware components with abnormal behavior [119] and replace them if necessary. To determine which components exhibit such abnormal behavior, the data acquired during operation and HWC are carefully evaluated by system experts. A dedicated LHC signal monitoring tool [120], simplifies the evaluation process through automatic data acquisition and threshold based fault

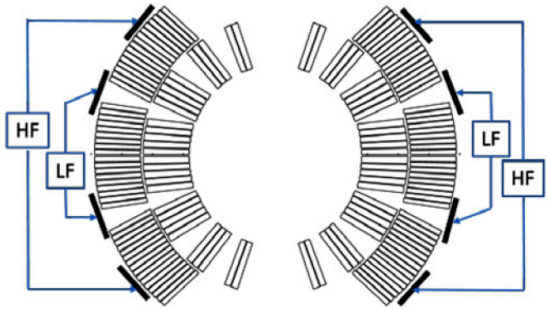
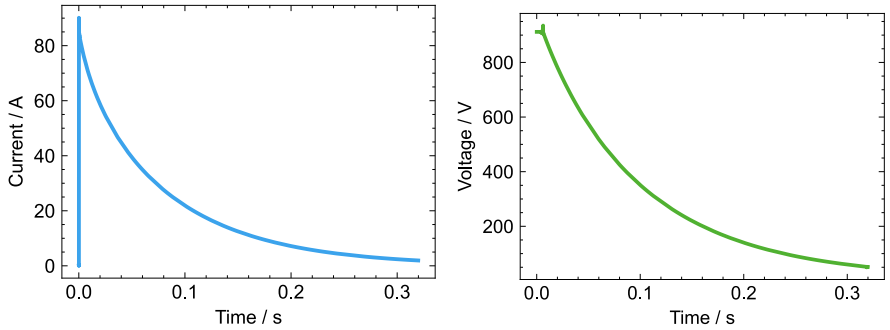


Figure 6.4: Cross-section of a magnet coil with two HF and two LF QHs attached to the outer surface [15].



(a) (b)

Figure 6.5: Voltage and the current measured at two resistive QH strips in the quenched magnet after a FPA event on 31.03.2021. The discharge is around 300 ms long and begins as soon as the FPA is triggered and the QHs are activated at 0 s.

prediction. Physical models further allow the experts to compare the data to simulations [109], [121] that replicate the normal and abnormal behavior of the magnet. In the next section, two new methods are presented which further support the experts in their data analysis.

6.2 Research Contribution 3

In the context of this research contribution, the frequency spectra of the voltages U_{M} measured at the main dipole magnets (see Fig. 6.3) are analyzed with unsupervised ML, as presented in Section 3.2.5. The results are presented in detail in Paper 6 and are summarized in Section 6.2.1. Furthermore, a method is presented to improve a current threshold based fault prediction method for QHs with supervised ML, as described in Section 3.2.3. The findings are shown in detail in Paper 1 and summarized in Section 6.2.2.

6.2.1 Anomaly Detection in the Main Dipole Magnets

With the current analysis tools used by experts, the frequencies occurring in the magnet voltage during the flat plateaus [0.2; 0.575] and [0.7; 1.075] seconds

after the triggering of the FPA (see Fig. 6.3) cannot be reconstructed. Therefore, these plateaus are investigated in more detail in this analysis. This is done by determining the frequency spectra $\mathbf{V}$ in the plateaus via a Fast Fourier Transformation (FFT), an efficient algorithm for computing the discrete Fourier transform [122]. If these frequencies are displayed for each magnet at which the voltage is measured and sorted by its electrical position, a Frequency Position Map (FPM) is derived. Such a FPM is shown in Fig. 6.6 for the magnet voltage in the plateau at [0.2; 0.575] seconds, visible in the magnified view of Fig. 6.3. During this event, a quench occurred at the electrical position 141 (white solid arrow in Fig. 6.6). The electrical positions 14 and 15 are the physical circuit neighbors of the quenched magnet (white empty arrow in Fig. 6.6). The brighter the color in the plot, the higher the amplitude of the frequency at the given electrical position. Different spectral components originate from various positions. These spectral components are examined in more detail below.

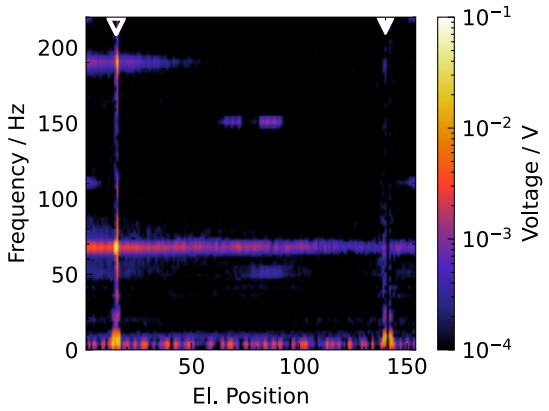


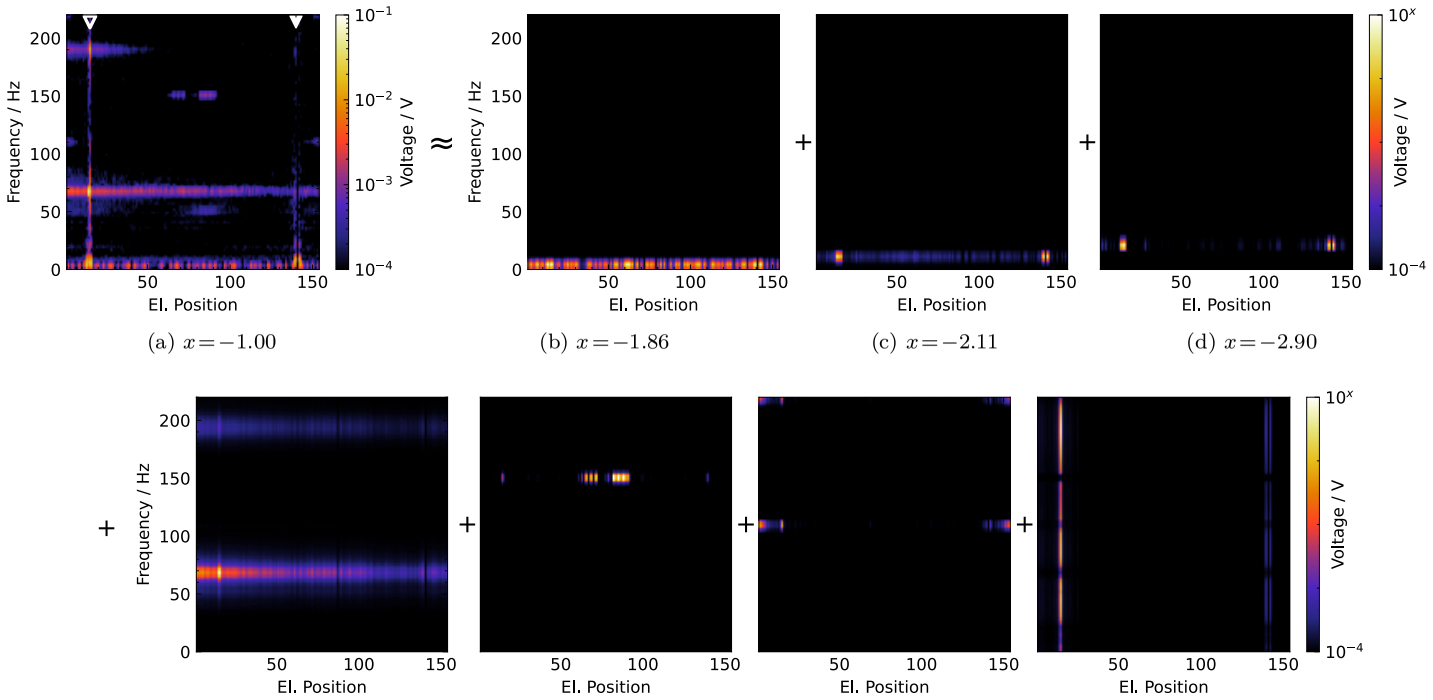
Figure 6.6: FPM of the frequencies occurring in the voltage signal, measured [0.2; 0.575] seconds after the triggering of the FPA of sector 78 on 31.03.2021. The solid white arrow marks the quenched magnet, while the empty arrow marks its physical neighbors.

Spectral Components

With the NMF presented in Section 3.2.5, the frequency spectra $\mathbf{V}$ are decomposed into spectral components $\mathbf{W}$ with weights $\mathbf{H}$. In total, $K = 7$ spectral components are identified. In Fig. 6.7, these spectral components are multiplied with the weights of the FPA event on 31.03.2021, to reconstruct the input FPM in Fig. 6.7a. A bright color represents a high amplitude. The maximum amplitude is 10^x V, where x is given in the caption of each figure. For better visibility, the frequency range is restricted to 0-220 Hz. The figures show the spectral components with amplitudes at (b) 3 Hz, (c) 6 Hz, (d) 20 Hz, (e) 66 Hz, (f) 150 Hz, and (g) 478 Hz. The last spectral component shows a (h) broadband spectrum.

Based on a statistical analysis of the weights $\mathbf{H}$, the physical processes behind each of the spectral components $\mathbf{W}$ are identified and summarized below using the example shown in Fig. 6.7:

- Spectral component one (SC1) is visible in Fig. 6.7b in the bright horizontal frequency band at 3 Hz. At positions 15 and 141, which are the



(a) $x = -1.00$ (b) $x = -1.86$ (c) $x = -2.11$ (d) $x = -2.90$

(e) $x = -1.52$ (f) $x = -2.85$ (g) $x = -3.22$ (h) $x = -3.02$

Figure 6.7: Frequency and amplitude of the identified seven spectral components as a function of the electrical position. (a) shows the input FPM with a quench in sector 78 on 31.03.2021. (b-h) show each spectral component multiplied with the weights of this FPA event, to reconstruct the initial FPM in (a). For better visibility, the maximum of the color axis is scaled with 10^x V. Additionally, the frequency range is restricted to 0-220 Hz, in which the majority of the spectral components occur.

physical and electrical neighbors of the quenched magnet, the spots are particularly bright. The physical process causing these high amplitudes are electromagnetic perturbations. They are induced by the quenched magnet to its physical and electrical neighbors. The remaining bright spots do not originate from a physical process, but are numerical artifacts introduced by the FFT.

- Spectral component two (SC2) is visible in Fig. 6.7c by two bright points at 6 Hz at the electrical positions around 15 and 141. Again, this spectral component represents electromagnetic perturbations induced by the quenched magnet.

- Spectral component three (SC3), illustrated in Fig. 6.7d, shows a similar pattern to SC2 and can be attributed to electromagnetic perturbations. In addition, SC3 is affected by the voltage waves traveling along the chain of magnets as governed by the magnet impedance [121].

- Spectral component four (SC4) is visible in Fig. 6.7e and shows a bright spot at 66 Hz and 184 Hz. The spots are the brightest at the magnet positions 14 and 15, which are the physical neighbors of the quenched magnet. From there, the oscillation is propagating along the electrical direction. While the exact physical process of SC4 remains elusive, it is expected that it is triggered by a quench.

- Spectral component five (SC5) appears as a double horizontal band at 150 Hz around the electrical position 77 in Fig. 6.7f. The bright spots of the band occur at exactly the same input of each measurement unit. SC5 only occurs in FPA events in sectors 12, 45, 67, 78, and 81.

- Spectral component six (SC6) is visible at 107 Hz and 220 Hz as spots originating from the electrical positions 1 and 154 of Fig. 6.7g. The magnets at these electrical positions are installed close to the power converters of the circuit. The amplitude of SC6 decreases with increasing distance from the power converter in sector 78. In addition, SC6 has high amplitudes at 260 Hz, 370 Hz, and 478 Hz, not visible in Fig. 6.7g due to the restricted range of the frequency axis. During the EE plateaus the PC is deactivated, indicating that SC6 originates from passive hardware components in the PC in sector 78.

- Spectral component seven (SC7), illustrated in Fig. 6.7h, shows one vertical line with high amplitude at the electrical positions 14 and 15 and one line with low amplitude at the electrical positions around 140 and 142. Both lines have interruptions at frequencies already reconstructed by other spectral components. These vertical lines indicate a broadband spectrum in magnets physically close to the quench. In the time domain, this broadband spectrum corresponds to a spike. In the QH discharge signals in Fig. 6.5, such spikes are used as indicators for intermittent short circuits in the resistive strips of the QHs [15]. Hence, SC7 might also be a critical indicator of an intermittent short circuit in the magnet.

Anomaly Detection

The described spectral components allow reconstructing normal FPA events with low reconstruction loss. If the reconstruction of a FPA event is not possible with a low loss, it is an abnormal FPA event. Across different hyperparameter combinations, four FPA events show a particular high reconstruction loss. The magnets that quenched during these FPA events are indicated with the LHC specific identifiers #2038, #1225, #1146, and #1291.

When considering the spectral components of these FPA events, also the amplitude of SC7 at the quenched magnet stands out. Only the FPA event with a quench in magnet #1146 is close to the average SC7 amplitude of 1 mV. For the FPA events where the magnets #2038, #1225, and #1291 quenched, the amplitudes of this spectral component are 240 mV, 80 mV, and 210 mV, respectively. These amplitudes are more than 80 times higher than the average, while they are not substantially elevated in normal events. It is inferred that a high SC7 in the quenched magnet is a strong indicator for identifying an anomaly. Based on this indicator, one additional FPA event was found that also showed an elevated SC7 amplitude of 1200 mV in the quenched magnet #2421, which is also referred to as an anomaly.

One of the four quenched magnets, with a significantly increased SC7 amplitude during a FPA event, has developed an intermittent short circuit during the FPA event on 25.04.2021. As such an intermittent short circuit is a critical event (see Fig. 6.1), the other three magnets #1225, #1291, and #2421 are also treated as potentially critical and will be checked by transient voltage measurement. If an intermittent short circuit cannot be excluded during the transient voltage measurement, these magnets could be replaced in one of the next maintenance stops of the LHC. In any case, the electronics of the measurement units of these magnets should be exchanged in order to exclude measurement errors.

Transient measurements will also be performed on the magnet #1146 as the amplitude of SC1 in the electrical and physical neighbors is particularly high in this event (1500 mV). These measurements will provide further information about electromagnetic perturbations.

Tab. 6.1 summarizes the five discussed anomalies, where the quenched magnet is stated in the first column. The second column shows the affected circuit and the date of the related FPA event. The last column summarizes main findings of the FPA event and states the recommended maintenance actions.

Table 6.1: List of detected anomalies with recommended maintenance actions in the remarks column.

Quenched Magnet	Abnormal FPA Event	Remarks
#2038	Sector 78 25.04.2021	High SC7 in #2038 (240 mV) Exchanged on 25.04.2021 due to intermittent short circuit
#1225	Sector 45 12.05.2021	High SC7 in #1225 (80 mV) Additional measurements Hardware replacements
#1146	Sector 34 06.05.2021	High SC1 (1500 mV) Additional measurements
#1291	Sector 12 14.05.2021	High SC7 in #1291 (210 mV) Additional measurements Hardware replacements
#2421	Sector 34 20.04.2021	High SC7 in #2421 (1200 mV) Additional measurements Hardware replacements

6.2.2 Fault Prediction in Quench Heaters

The prediction of QH faults is crucial for a timely activation of the spare LF QHs to ensure continuous magnet protection. However, a QH fault can also damage the magnet itself. Therefore, the time-series signals $\boldsymbol{X}_n$ shown in Fig. 6.5, are monitored with a Quench Heater Discharge Analysis (QHDA) tool.

The approach with the QHDA tool is shown in Fig. 6.8. Four features $\boldsymbol{x}_n^\star$ are calculated with a function $\phi_f(\cdot)$. These features are [15]:

1. *Steady state voltage level:* The voltage values at the start and end of the QH discharge in Fig. 6.5b are used.

2. *Characteristic time of the pseudo-exponential decay:* The characteristic time of the pseudo-exponential decay is determined from the voltage and the current signals during the QH discharge.

3. *Steady state resistance level:* The resistance of the QH strips is determined from the voltage and current signals after the activation of the QH discharge.

4. *Signal comparison:* The signals and the features 1. to 3. are compared with the sample-wise L1-norm to reference discharges of the corresponding QH circuit.

With the features $\boldsymbol{x}_n$, the tool then predicts a label healthy ($\hat{y}_n = 1$) or faulty ($\hat{y}_n = -1$) for the event with a threshold function $g(\cdot)$. System experts set these thresholds, and check faulty predictions manually. The true label is

indicated by y_n. When the prediction is false negative ($\hat{y}_n = -1, y_n = 1$), $\phi_f(\cdot)$ and $g(\cdot)$ are adjusted.

In case of a true negative prediction ($\hat{y}_n = y_n = -1$), the cause of the fault is investigated in detail. A fault can occur in the form of a short circuit between QH strip to magnet coil, QH strip to ground, and QH strip to magnet coil and ground. In the first two scenarios, the activation of the spare LF QHs is often sufficient. However, in the latter case, the operation must be stopped immediately, and the magnet has to be replaced in a long and tedious process.

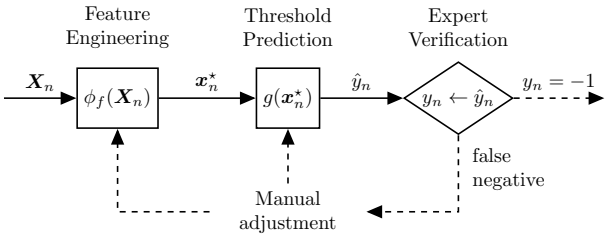


Figure 6.8: Approach with the QHDA: The calculation of the features is followed by a threshold based fault prediction. This prediction is checked by system experts to find faulty QHs with label $y_n = -1$.

Healthy predictions are not checked by experts. Therefore, a false positive prediction ($\hat{y}_n = 1, y_n = -1$) only emerges if other protection systems are triggered or if damage occurs. To minimize the number of false positive predictions, the QHDA tool is extended with ML to derive a hybrid fault prediction.

Hybrid Fault Prediction

The approach with the hybrid fault prediction is shown in Fig. 6.9. Both the QHDA tool and the SVM use the same features $\boldsymbol{x}_n^\star$ as input. In the QHDA tool, the thresholds of $g(\cdot)$ are set by experts, while the SVM optimizes the decision boundary $h(\cdot)$ as described in Section 3.2.3. The output $\hat{y}_n^{\text{thb}}$ of the QHDA tool

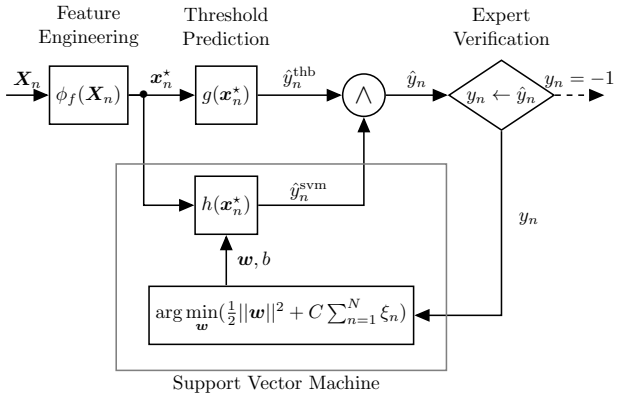


Figure 6.9: Approach with the hybrid fault prediction: The QHDA and the SVM use the same features $\boldsymbol{x}_n^\star$ to jointly predict the label $\hat{y}_n$. The SVM learns from the past decisions of the expert, by optimizing the parameter $\boldsymbol{w}$ in the decision boundary $h(\boldsymbol{x}_n)$.

is connected with the output $\hat{y}_n^{\text{svm}}$ of the SVM, such that

$$\hat{y}_n = \begin{cases} 1 & \text{if } \hat{y}_n^{\text{thb}} = 1 \wedge \hat{y}_n^{\text{svm}} = 1 \\ -1 & \text{otherwise.} \end{cases} \tag{6.1}$$

Once the experts have validated the prediction, the label y_n of the event is added to the dataset and the SVM parameters $\boldsymbol{w}$ and b are recalculated by optimizing $\arg\min_{\boldsymbol{w}}(\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{n=1}^{N}\xi_n)$, as described in Eq. (3.4). The adjustment of $g(\cdot)$, carried out manually in the previous approach, therefore takes place automatically.

The approach is validated with 3130 healthy and 116 faulty events, recorded during the whole second operational run of the LHC between 2015 and 2018. The prediction results are compared to the approach with the QHDA tool in Tab. 6.2. The True Positive (TP) rate indicates the ratio of correctly predicted healthy events, while the False Positive (FP) rate indicates the ratio of false healthy predictions. Although the SVM FP rate of 13.1% is similar to the QHDA tool FP rate of 14.7%, the combination of the two prediction models yields a FP rate of 2.4%. This corresponds to 14 additional events that are correctly predicted as faults. The TP rate of the hybrid model is 1% lower compared to the QHDA model, which corresponds to around 31 events. Therefore, experts would have checked 45 additional QH discharge events. For 14 events, they would have taken additional maintenance actions to minimize the risk of a short to ground of a QH strip to the magnet coil and the ground.

Table 6.2: True Positive (TP) rate and False Positive (FP) rate of the fault prediction in QHs during the second operational run of the LHC.

Method	TP rate	FP rate
QHDA tool	99.9%	14.7%
SVM model	99.1%	13.1%
Hybrid approach	98.9%	2.4%

Chapter 7

Conclusion

Machine Learning (ML) based fault prediction offers an opportunity to meet the ambitious availability targets of CERN's energy frontier colliders. However, the limited number of faults of a single type makes it challenging to validate the results with statistical tests to ensure that the predictions are reliable. In this thesis, three approaches were successfully studied to overcome this limitation. Those approaches are based on hybrid models, Explainable Artificial Intelligence (XAI), and interpretable ML models, summarized below.

Hybrid Models

Existing tools for fault prediction can be further improved with ML. This was shown with a hybrid model, which connects the prediction of the existing Quench Heater Discharge Analysis (QHDA) tool and a non-linear Support Vector Machine (SVM) with a logical OR. The method was tested on data of the second operational run of the Large Hadron Collider (LHC). In this time period, the risk of an intermittent short circuit in a magnet via the faulty quench heaters was reduced by 12.3%.

Explainable Artificial Intelligence

XAI allows identifying the fault indicator of ML predictions to validate their credibility. A novel model-agnostic XAI method for explaining fault predictions to system experts was presented. The autoencoder method generates concept-based prototype explanations for pre-trained supervised ML models. In a test, the explanations of this method helped 73 system experts to identify fault indicators at least 5.5% better than two similar XAI methods.

This XAI method was then applied to interpret follow-up breakdown predictions in Radio Frequency (RF) cavities of the Compact Linear Collider (CLIC). With ML-based breakdown prediction, the negative consequences of a follow-up breakdown could have been mitigated in 89.7% of all cases. The XAI method revealed the likely source of this prediction: Fluctuations in the electron emissions after an initial breakdown increase the probability of another breakdown occurring shortly after.

Interpretable Models

With the interpretable method Non-negative Matrix Factorization (NMF), the voltages measured across the 1232 superconducting dipole magnets in the eight LHC main dipole circuits were analyzed to understand the normal and abnormal behavior of the circuits. This allowed the extraction of seven spectral components that define normal behavior, occurring in the measured voltages during a Fast Power Abort (FPA) event. Analyzing the spectral components' distribution and propagation across the circuit and across FPA events provided a deeper understanding of the mutual interaction of hardware components and allowed identifying the potential physical processes causing the spectral components.

Five additional magnets with abnormal behavior during FPA events were detected, using the reconstruction loss of NMF and the amplitude of a single spectral component, which was emphasized in all outliers. One of these magnets was replaced after a short circuit was detected on 25.04.2021 following a FPA event. This is a strong indication for the validity of the method. Similarly to the replaced magnet, three of the four remaining magnets showed an elevated amplitude in one spectral component during their quench, which was more than 80 times higher than normal. The three magnets could be replaced in one of the next maintenance stops of the LHC to prevent up to three months of unplanned LHC downtime.

Overall, it has been demonstrated that interpretable ML is a powerful tool to identify faults in a highly complex and specialized system like the LHC, avoiding long and expensive downtime of this unique facility. Applying these methods on a large scale in future energy frontier colliders is essential to achieve their very high and ambitions availability targets.

Bibliography

[1] O. S. Brüning, P. Collier, P. Lebrun, *et al.*, *LHC Design Report*. CERN, 2004.

[2] R. Vanden Broeck, *Beam Dump - Absorbeur de faisceau*, CERN Document Server, 2019.

[3] M. Aicheler *et al.*, *The Compact Linear Collider (CLIC) – Project Implementation Plan*. CERN, 2018.

[4] M. Benedikt, A. Blondel, O. Brunner, *et al.*, *FCC-ee: The Lepton Collider: Future Circular Collider Conceptual Design Report Volume 2.* CERN, 2019.

[5] M. Benedikt, M. Capeans Garrido, F. Cerutti, *et al.*, *FCC-hh: The Hadron Collider: Future Circular Collider Conceptual Design Report Volume 3.* CERN, 2019.

[6] W. Wuensch, N. Catalán Lasheras, A. Degiovanni, *et al.*, "Experience operating an x-band high-power test stand at CERN," in *Proceedings of the 5th International Particle Accelerator Conference*, Dresden, Germany, 2014, pp. 2288–2290.

[7] J. Steckert, R. Denz, S. Mundra, T. Podzorny, J. Spasic, and D.-G. Vancea, "Application of the universal quench detection system to the protection of the high-luminosity LHC magnets at CERN," *Transactions on Applied Superconductivity*, vol. 32, no. 6, pp. 1–5, 2022.

[8] C. Tennant, A. Carpenter, T. Powers, A. Shabalina Solopova, L. Vidyaratne, and K. Iftekharuddin, "Superconducting radio-frequency cavity fault classification using machine learning at jefferson laboratory," *Phys. Rev. Accel. Beams*, vol. 23, no. 11, pp. 114 601–114 617, 2020.

[9] D. Hoang, C. Boffo, N. Tran, *et al.*, "Intelliquench: An adaptive machine learning system for detection of superconducting magnet quenches," *Transactions on Applied Superconductivity*, vol. 31, no. 5, pp. 1–5, 2021.

[10] J. Beítez, J. Castro, and I. Requena, "Are artificial neural networks black boxes?" *Transactions on Neural Networks*, vol. 8, no. 5, pp. 1156–1164, 1997.

[11] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation'," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[12] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, "How can i explain this to you? an empirical study of deep neural network explanation methods," in *Advances in Neural Information Processing Systems*, Online Conference, 2020, pp. 4211–4222.

[13] J. Paszkiewicz, P. Burrows, and W. Wuensch, "Spatially resolved dark current in high gradient traveling wave structures," in *Proceedings of the 10th International Particle Accelerator Conference*, Geneva, Switzerland, 2019.

[14] E. Z. Engelberg, J. Paszkiewicz, R. Peacock, S. Lachmann, Y. Ashkenazy, and W. Wuensch, "Dark current spikes as an indicator of mobile dislocation dynamics under intense dc electric fields," *Physical Review Accelerators and Beams*, vol. 23, no. 12, p. 123 501, 2020.

[15] Z. Charifoulline, L. Bortot, R. Denz, *et al.*, "Overview of the performance of quench heaters for high-current LHC superconducting magnets," *IEEE Transactions on Applied Superconductivity*, vol. 27, no. 4, pp. 1–5, 2016.

[16] L. Bortot, B. Auchmann, I. Garcia, *et al.*, "STEAM: A Hierarchical Cosimulation Framework for Superconducting Accelerator Magnet Circuits.," *Transactions on Applied Superconductivity*, vol. 28, no. 3, p. 4 900 706, 2017.

[17] T. K. Kroc, J. C. Thangaraj, R. T. Penning, and R. D. Kephart, "Accelerator-driven medical sterilization to replace Co-60 sources," Fermi National Accelerator Lab., Fermilab Pub. 17-314-DI, 2017.

[18] U. Amaldi, "Cancer therapy with particle accelerators," *Nuclear Physics A*, vol. 654, no. 1-2, pp. C375–C399, 1999.

[19] G. Cuttone, "Applications of particle accelerators in medical physics," CERN, Knowledge Transfer 2013-001, 2008.

[20] M. Jensen, "Particle accelerators for pet radionuclides," *Nuclear Medicine Review*, vol. 15, no. C, pp. 9–12, 2012.

[21] R. W. Hamm and M. E. Hamm, "The beam business: Accelerators in industry," *Physics Today*, vol. 64, no. 6, pp. 46–51, 2011.

[22] M. Current, L. Rubin, F. Sinclair, and J. Ziegler, *Commercial ion implantation systems*. Ion Implant Technology, 2018.

[23] C. Tang, "Low energy accelerators for cargo inspection," *Reviews of Accelerator Science and Technology*, vol. 8, no. 1, pp. 143–163, 2015.

[24] F. Lucarelli, "How a small accelerator can be useful for interdisciplinary applications: The study of air pollution," *The European Physical Journal Plus*, vol. 135, no. 7, p. 538, 2020.

[25] CERN, *Who we are: Our people*, `https://home.cern/about/who-we-are/our-people`, Accessed: 23.10.2023.

[26] ATLAS Collaboration, "Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC," *Phys. Lett. B*, vol. 716, pp. 1–29, 2012. arXiv: `1207.7214`.

[27] C. Adolphsen, D. Angal-Kalinin, T. Arndt, *et al.*, "European strategy for particle physics–accelerator R&D roadmap," *CoRR*, 2022. arXiv: `2201.07895`.

[28] J. A. Osborne, A. Tudora, M. J. Stuart, and E. Ter Laak, "FCC and CLIC layout," 2019.

[29] M. Bajko, L. Rossi, R. Schmidt, *et al.*, "Report of the task force on the incident of 19th september 2008 at the LHC," CERN, LHC Project Report 1168, 2009.

[30] K. Foraz *et al.*, "LS1, first long shutdown of LHC and its injector chains," CERN, Tech. Rep. 2014-0223, 2014.

[31] J.-P. Tock, M. Bednarek, L. Bottura, *et al.*, "The second LHC long shutdown (LS2) for the superconducting magnets," in *Proceedings of the 9th International Particle Accelerator Conference*, Vancouver, Canada, 2018, p. 39.

[32] CERN, *LHC long term schedule*, `https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm`, Accessed: 23.10.2023.

[33] A. G. *et al.*, *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1.* CERN, 2017.

[34] K. Hanke *et al.*, "The LHC injectorsupgrade (LIU) project at CERN: Proton injector chain," in *Proceedings of the 8th International Particle Accelerator Conference*, Copenhagen, Denmark, 2017, pp. 3335–3338.

[35] G. Dissertori, "LHC detectors and early physics," *Proceedings of the 65th Scottish Universities Summer School in Physics*, pp. 1–197, 2012.

[36] E. Lopienska, "The CERN accelerator complex, layout in 2022," 2022.

[37] P. G. Huray, *Maxwell's Equations*. John Wiley & Sons, 2009.

[38] D. Boussard and T. P. R. Linnecar, "The LHC superconducting RF system," CERN, LHC Project Report 316, 1999.

[39] R.-L. Geng, H. Padamsee, A. Seaman, and V. D. Shemelin, "World record accelerating gradient achieved in a superconducting niobium rf cavity," in *Proceedings of the 2005 Particle Accelerator Conference*, IEEE, Tennessee, USA, 2005, pp. 653–655.

[40] CLIC, *Clic - home*, `https://clic.cern`, Accessed: 23.10.2023.

[41] A. Magazinik, N. Catalan-Lasheras, and J. Sauza Bedolla, "Industrialization study for 12 GHz accelerating structures for CLIC 380," CERN, CLIC Note 1178, 2021.

[42] O. Brunner, P. N. Burrows, S. Calatroni, *et al.*, *The CLIC project*, CERN, 2022. arXiv: `2203.09186`.

[43] M. Mangano, P. Azzi, M. Benedikt, *et al.*, *FCC Physics Opportunities: Future Circular Collider Conceptual Design Report Volume 1. Future Circular Collider*. CERN, 2019.

[44] E. Wilson and B. J. Holzer, *Particle Physics Reference Library: Volume 3: Accelerators and Colliders*. Springer Nature, 2020.

[45] B. Humann, F. Cerutti, and R. Kersevan, "Synchrotron radiation impact on the FCC-ee arcs," in *Proceedings of the 12th International Particle Accelerator Conference*, Bangkok, Thailand, 2022, pp. 1675–1678.

[46] T. Raubenheimer, *FCC accelerator overview*, Presentation at the FCC Week Workshop, FCC collaboration & FCCIS DS team, 2023.

[47] B. Goddard *et al.*, "Main changes to LHC layout for reuse as FCC-hh high energy booster," CERN, Tech. Rep. CERN-ACC-2015-0030, 2015.

[48] O. Rey Orozco, *Availability of particle accelerators: requirements, prediction methods and optimization*. Stuttgart: Institut für Maschinenelemente, 2020.

[49] M. S. Camillocci, "LHC nominal cycle," in *6th Evian Workshop on LHC beam operation*, Geneva, Switzerland: CERN, 2016, pp. 45–48.

[50] C. "Roderick, L. Burdzanowski, D. Martin Anido, S. Pade, and P. Wilk, "Accelerator fault tracking at CERN," in *16th International Conference on Accelerator and Large Experimental Physics Control Systems*, Barcelona, Spain, 2018, pp. 397–400.

[51] J. Heron, *The Availability Challenge: Targets, shortfalls and game-changing opportunities*, Presentation at the FCC Week Workshop, CERN, 2023.

[52] F. Bordry, M. Benedikt, O. Bruning, *et al.*, "Machine Parameters and Projected Luminosity Performance of Proposed Future Colliders at CERN," CERN, Geneva, Tech. Rep. 2018. arXiv: 1810.13022.

[53] J. Cai, I. Syratchev, and G. Burt, *Two stage high efficiency klystron for FCC-ee*, Presentation at FCC Week 2021, 2021.

[54] L. Felsberger, A. Apollonio, T. Cartier-Michaud, E. Montesinos, J. Oliveira, and J. Uythoven, "Availability modeling of the solid-state power amplifiers for the CERN SPS RF upgrade," in *Proceedings of the 12th International Particle Accelerator Conference*, Campinas, Brazil, 2021, pp. 2308–2311.

[55] M. D. Castro *et al.*, *Robotic solutions for the inspection and remote maintenance of particle accelerators*, Presentation at FCC Week 2023, 2023.

[56] C. Wiesner, C. Hernalsteens, F. Hülphers, P. Ziegler, J. Uythoven, and D. Wollmann, "Automated evaluation of LHC proton losses during high-energy beam dumps for the post-mortem system," in *Proceedings of the 14th International Particle Accelerator Conference*, Venice, Italy, 2023, pp. 614–617.

[57] R. J. Rossi, *Mathematical Statistics: An Introduction to Likelihood Based Inference*. John Wiley & Sons, 2018.

[58] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006.

[59] N. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 1, pp. 175–185, 1992.

[60] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[61] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 1, pp. 273–297, 1995.

[62] E. Fol, R. Tomás, J. Coello de Portugal, and G. Franchetti, "Detection of faulty beam position monitors using unsupervised learning," *Phys. Rev. Accel. Beams*, vol. 23, no. 10, pp. 102 805–102 815, 2020.

[63] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[64] C. Emma, A. Edelen, M. J. Hogan, B. O'Shea, G. White, and V. Yakimenko, "Machine learning-based longitudinal phase space prediction of particle accelerators," *Phys. Rev. Accel. Beams*, vol. 21, no. 11, pp. 112 802–112 808, 2018.

[65] P. Arpaia, G. Azzopardi, F. Blanc, *et al.*, "Machine learning for beam dynamics studies at the CERN Large Hadron Collider," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 985, no. 1, p. 164 652, 2021.

[66] A. Pollard, A. Gilfellon, and D. Dunning, "Machine learning for rf breakdown detection at CLARA," in *Proceedings of the Linear Acceleration Conference*, Shanghai, China, 2022, pp. 681–685.

[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[68] D. D. Lee and H. S. Seung, "Learning the parts of speech by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[69] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, pp. 1–39, 2012.

[70] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," vol. 96, no. 34, pp. 226–231, 1996.

[71] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings 5th Berkeley Symp. Math. Statist. Prob.*, California, USA, 1967, pp. 281–297.

[72] D. Turner, R. Bachimanchi, A. Carpenter, J. Latshaw, C. Tennant, and L. Vidyaratne, "SRF cavity instability detection with machine learning at CEBAF," in *5th North American Particle Accel. Conf.*, Albuquerque, USA, 2022, pp. 669–671.

[73] J. P. Edelen and N. M. Cook, "Anomaly detection in particle accelerators using autoencoders," *CoRR*, 2021. arXiv: `2112.07793`.

[74] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors," in *International conference on ML*, Vienna, Austria, 2018, pp. 2668–2677.

[75] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," *CoRR*, 2016. arXiv: `1602.04938`.

[76] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, Long Beach, USA, 2017, pp. 4768–4777.

[77] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, pp. 1–46, 2015.

[78] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Interpretable deep learning by propagating activation differences," *CoRR*, 2016. arXiv: 1605.01713.

[79] K. Simonyan, A. Vedaldi, and A. Zisserman, "Visualising image classification models and saliency maps," *CoRR*, 2013. arXiv: 1312.6034.

[80] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Winter conference on applications of computer vision*, Lake Tahoe, USA, 2018, pp. 839–847.

[81] L. S. Shapley, *A value for n-person games*. Princeton University Press, 1988.

[82] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Advances in Neural Information Processing Systems*, Long Beach, USA, 2017, pp. 6949–6958.

[83] R. Mochaourab, A. Venkitaraman, I. Samsten, P. Papapetrou, and C. R. Rojas, "Post-hoc explainability for time series classification: Toward a signal processing perspective," *signal processing magazine*, 2022.

[84] C.-K. Yeh *et al.*, "On completeness-aware concept-based explanations in deep neural networks," in *Advances in Neural Information Processing Systems*, Online Conference, 2020, pp. 20 554–20 565.

[85] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes," *CoRR*, 2017. arXiv: 1710.04806.

[86] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," *CoRR*, 2019. arXiv: 1904.0.

[87] R. Guidotti, A. Monreale, F. Spinnato, D. Pedreschi, and F. Giannotti, "Explaining any time series classifier," in *Second International Conference on Cognitive Machine Intelligence*, Atlanta, USA, 2020, pp. 167–176.

[88] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 8930–8941.

[89] S. Das, P. Xu, Z. Dai, A. Endert, and L. Ren, "Interpreting deep neural networks through prototype factorization," in *International Conference on Data Mining Workshops*, IEEE, Sorrento, Italy, 2020, pp. 448–457.

[90] W. Tang, L. Liu, and G. Long, "Interpretable time-series classification on few-shot samples," in *International Joint Conference on Neural Networks*, Glasgow, UK, 2020, pp. 1–8.

[91] C.-K. Yeh *et al.*, "On completeness-aware concept-based explanations in deep neural networks," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 20 554–20 565.

[92] H. A. Dau, A. Bagnall, K. Kamgar, *et al.*, "The UCR time series archive," *Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.

[93] R. T. Olszewski, "Generalized feature extraction for structural pattern recognition in time series data," Ph.D. dissertation, 2001.

[94] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical Science*, vol. 16, no. 2, pp. 101–133, 2001.

[95] T. Lucas, "High field phenomenology in linear accelerators for the compact linear collider," Ph.D. dissertation, PhD Thesis, University of Melbourne, 2018. from this w ork may be used under ..., 2018.

[96] B. Woolley, G. Burt, A. C. Dexter, *et al.*, "High-gradient behavior of a dipole-mode rf structure," *Phys. Rev. Accel. Beams*, vol. 23, no. 12, p. 122 002, 2020.

[97] T. G. Lucas, T. Argyopolous, M. J. Boland, *et al.*, "Dependency of the capture of field emitted electron on the phase velocity of a high-frequency accelerating structure," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 914, no. 1, pp. 46–52, 2019.

[98] R. Rajamaki, "Vacuum arc localization in clic prototype radio frequency accelerating structures," Master's thesis, Aalto University, 2016.

[99] W. Wuensch, A. Degiovanni, S. Calatroni, *et al.*, "Statistics of vacuum breakdown in the high-gradient and low-rate regime," *Phys. Rev. Accel. Beams*, vol. 20, no. 1, pp. 011 007–011 018, 2017.

[100] W. L. Millar, A. Grudiev, W. Wuensch, *et al.*, "High-power test of two prototype X-band accelerating structures based on SwissFEL fabrication technology," *Transactions on Nuclear Science*, vol. 70, no. 1, pp. 1–19, 2023.

[101] B. Woolley, "High power X-band RF test stand development and high power testing of the CLIC crab cavity," Ph.D. dissertation, Lancaster University, 2015.

[102] R. Wegner, *Strategy for RF breakdown protection of the Linac4 RFQ*, Presentation at the 188th Machine Protection Panel Meeting (Injectors): Special meeting on Linac4 RFQ protection, CERN, 2020.

[103] L. S. Cowie, *Normal conducting RF structure development for CLARA*. Lancaster University, 2021.

[104] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[105] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[106] E. Z. Engelberg, Y. Ashkenazy, and M. Assaf, "Stochastic model of breakdown nucleation under intense electric fields," *Phys. Rev. Lett.*, vol. 120, no. 12, p. 124 801, 2018.

[107] J. Paszkiewicz, "Studies of breakdown and pre-breakdown phenomena in high-gradient accelerating structures," Ph.D. dissertation, St. John's College, Oxford, 2021.

[108] L. Rossi and E. Todesco, "Conceptual design of 20 T dipoles for high-energy LHC," *CoRR*, 2011. arXiv: 1108.1619.

[109] E. Ravaioli, K. Dahlerup-Petersen, F. Formenti, *et al.*, "Impact of the voltage transients after a fast power abort on the quench detection system in the LHC main dipole chain," *Transactions on Applied Superconductivity*, vol. 22, no. 3, p. 9 002 504, 2012.

[110] J. Schultz, "Protection of superconducting magnets," *Transactions on Applied Superconductivity*, vol. 12, no. 1, pp. 1390–1395, 2002.

[111] R. Denz and F. Rodriguez-Mateos, "Electronic systems for the protection of superconducting elements in the LHC," *Transactions on Applied Superconductivity*, vol. 16, no. 2, pp. 1725–1728, 2006.

[112] R. Denz, K. Dahlerup-Petersen, F. Formenti, *et al.*, "Upgrade of the protection system for superconducting circuits in the LHC," in *Proceedings of the 11th Particle Accelerator Conference*, Genoa, Italy, 2009, pp. 244–246.

[113] F. Rodriguez-Mateos, P. Pugnat, S. Sanfilippo, R. Schmidt, A. Siemko, and F. Sonnemann, "Quench heater experiments on the LHC main superconducting magnets," in *Proceedings of the 7th Particle Accelerator Conference*, Vienna, Austria, 2000, pp. 2154–2156.

[114] F. Sonnemann, "Resistive transition and protection of LHC superconducting cables and magnets," Ph.D. dissertation, RWTH Aachen University, 2001.

[115] F. Bourgeois and K. Dahlerup-Petersen, "Methods and results of modeling and transmission-line calculations of the superconducting dipole chains of CERN's LHC collider," in *28th IEEE International Conference on Plasma Science and 13th IEEE International Pulsed Power Conference*, 2001, pp. 804–807.

[116] K. Dahlerup-Petersen, F. Rodriguez-Mateos, R. Schmidt, and F. Sonnemann, "Energy extraction for the LHC superconducting circuits," in *Proceedings of the 5th Particle Accelerator Conference*, Chicago, USA, 2002, pp. 3448–3450.

[117] E. Ravaioli, K. Dahlerup-Petersen, F. Formenti, J. Steckert, H. Thiesen, and A. Verweij, "Modeling of the voltage waves in the LHC main dipole circuits," *Transactions on Applied Superconductivity*, vol. 22, no. 3, p. 9 002 704, 2012.

[118] A. Apollonio, O. Andreassen, A. Antoine, T. Argyropoulos, M. Bastos, M. Bednarek, *et al.*, "Summary of the post-long shutdown 2 LHC hardware commissioning campaign," in *Proceedings of the 13th International Particle Accelerator Conference*, Bangkok, Thailand, 2022, pp. 335–338.

[119] R. Saederup, "Local transfer function measurement data analysis," CERN, Tech. Rep. 2675917, 2021.

[120] C. Obermair, "Signal monitoring for the LHC," pp. 1–10, 2018.

[121] M. Janitschke, "Framework for automatic superconducting magnet model generation & validation against transients measured in LHC magnets," M.S. thesis, Technical University of Berlin, 2021.

[122] H. J. Nussbaumer, *The Fast Fourier Transform*. Springer, 1981.

Part II

Included Papers

Paper 1

Machine Learning with a Hybrid Model for Monitoring of the Protection Systems of the LHC

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB345

MACHINE LEARNING WITH A HYBRID MODEL FOR MONITORING OF THE PROTECTION SYSTEMS OF THE LHC

C. Obermair*[1], M. Maciejewski, Z. Charifoulline, A. Apollonio, A. Verweij
CERN, Geneva, Switzerland
F. Pernkopf, Graz University of Technology, Graz, Austria
[1]also at Graz University of Technology, Graz, Austria

Abstract

The Large Hadron Collider (LHC) is the world's largest particle accelerator and uses a complex set of sophisticated and highly reliable machine protection systems to ensure a safe operation with high availability for particle physics production. The data gathered during several years of successful operation allow the use of data-driven methods to assist experts in finding anomalies in the behavior of those protection systems. In this paper, we derive a model that can extend the existing signal monitoring applications for the LHC protection systems with machine learning. Our hybrid model combines an existing threshold-based system with a Support Vector Machine (SVM) by using signals, manually validated by experts. Even with a limited amount of data, the SVM learns to integrate the expert knowledge and contributes to a better classification of safety critical signals. Using this approach, we analyze historical signals of quench heaters, which are an important part of the quench protection system for superconducting magnets. Particularly, it is possible to incorporate expert decisions into the classification process and to improve the failure detection rate of the existing quench heater discharge analysis tool.

INTRODUCTION

The early detection of faulty components contributes significantly to increasing machine availability and, thus, the LHC's physics potential expressed in terms of integrated luminosity. To protect the highly critical systems the machine protection system ensures safe operation of accelerator equipment (e.g. the superconducting magnets) and protects it from damage [1]. If the system fails it can result in an LHC downtime in the order of three months. Therefore, it requires consistent supervision of the components through signal monitoring and regular hardware commissioning tests. The quench protection system is part of the machine protection system and the Quench Heaters (QHs) are an essential part of it. The purpose of the QHs is to expand the quenching region of a superconductor, in order to enlarge the area of energy dissipation and, thus, reduce the potentially dangerous hot spot temperatures in the superconducting material. All of the 1232 LHC main dipole magnets are equipped with eight QH circuits. During magnet operation four out of the eight QH circuits are ready to be operated in case of a magnet quench. The other four QH circuits provide redundancy, in case of a fault in one of the other four circuits.

* christoph.obermair@cern.ch

Existing signal monitoring applications are based on the calculation of characteristic features representing a signal, which are compared to fixed thresholds. These thresholds give a clear answer whether the signal is healthy or faulty. This approach is particularly effective because experts can incorporate their knowledge about the behavior of a component's degradation into the analysis process. The quench heater discharge analysis tool [2] is one application which makes use of this, but several other components of the LHC machine protection system use the same approach [3–5].

Since the first commissioning of the LHC in 2008, the amount of system supervision data is growing and alternative signal monitoring approaches such as machine learning are currently gaining attention [6]. Several efforts have been made in the past to show the potential of machine learning for LHC protection systems, e.g. to observe anomalous behaviors of LHC superconducting magnets [7]. Another approach [2] used a feed forward neural network to analyze quench heater discharges. However, the lack of "faulty" signals often prevents machine learning models to reach the necessary reliability to replace existing analysis tools [2].

In other fields of research [8–10] it is common to build ensembles of different classifiers in order to make use of the so called *wisdom of the crowd* [11], which allows classifiers to contribute to a better overall classification result (e.g. XGBoost [12]). However, a hybrid approach which combines the advantages of existing LHC signal monitoring applications with the advantages of machine learning models has not been considered yet. Thus, the objective of this work was to develop an approach that allows LHC signal monitoring applications to benefit from the growing amount of historical data.

The paper is structured as follows: First the concept of threshold-based signal monitoring applications is explained and the workflow of the hybrid classification approach is derived. Subsequently, the existing analysis of past quench heater discharges is presented and the results of the new hybrid classification approach are discussed. Finally, the strengths and limitations of the approach are discussed and the conclusion is presented.

DEVELOPMENT OF THE MODEL

Typically, an LHC signal monitoring application processes windows of C multivariate discrete signals. Those data batches are represented by $Z_d \in \mathbb{R}^{N \times C}$, where N is the amount of samples and $d \in [1, ..., D]$ is the window index. Depending on the use case, such a batch can either have a

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB345

fixed or variable length and the beginning is either defined by a specific event, like a quench, or a manually chosen event, like a particular state of operation represented by the so-called beam mode.

Threshold-Based Classification

The workflow of a threshold-based signal monitoring approach is summarized in Fig. 1. Formally this means, that out of each signal batch matrix Z_d a function ϕ_f calculates F features $x_d \in \mathbb{R}^F$.With those features x_d, the application then assigns a label "healthy" ($y_d = 1$) or "faulty" ($y_d = -1$) to each batch with a threshold function:

$$g(x_d) = \begin{cases} 1 & \text{if } \check{k} < x_d < \hat{k} \\ -1 & \text{otherwise,} \end{cases} \tag{1}$$

in which $\check{k}$ is the minimum threshold vector and $\hat{k}$ is the maximum threshold vector, both determined by experts.

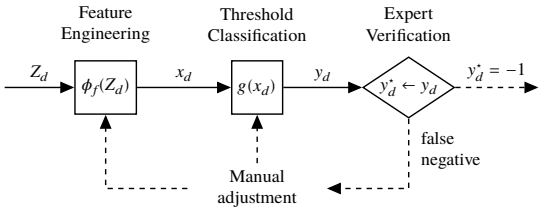


Figure 1: Workflow of a threshold-based signal monitoring approach. First the features are calculated, then a threshold is set to assign a label to the signal. This label is validated by experts in the last step.

In case the signal condition is identified as "faulty", experts have to verify this result. Consequently, if the experts decide that the prediction of the classification algorithm was true negative ($y_d^\star = -1$), they can initiate further actions, like a hardware inspection. However, if the experts decide that the automatic signal classification does not reflect the actual condition of a component (false negative), the machine operation continues as usual. Furthermore, the experts could then adjust the classification algorithm and/or the thresholds, such that this specific "faulty" classification does not occur in the future. However, due to the high amount of signals, experts often only get notified in case of a "faulty" classification. This means they can only intervene if the classification was false negative. A false positive label, which is a "faulty" signal labeled as "healthy", only emerges if other protection systems are triggered or if damage occurs.

Machine Learning Based Classification

During classification with machine learning models the parameters of a threshold function are optimized such that the best classification on a given input data set is reached. Common classification algorithms include logistic regression, random forest, neural networks, and SVMs [13]. We use the latter as it is especially suited for handling data sets with limited amount of samples and high dimensions.

The workflow of a machine learning based classification is similar to the threshold-based classification, but the threshold function is defined by the separation hyperplane:

$$h(x_d) = w^{\mathrm{T}} \phi(x_d) + b, \tag{2}$$

where w contains the weight parameters, b is the bias parameter and $\phi(x_d)$ is a fixed feature space transformation. Those parameters are determined by solving the following optimization problem, using training data, i.e.

$$\begin{aligned} \arg\min_{w} \quad & (\frac{1}{2}\|w\|^2 + C \sum_{d=1}^{D} \xi_d) \\ \text{subject to: } & y_d^\star (w^{\mathrm{T}} x_d + b) \geq 1 - \xi_d, \quad d = 1, ..., D \\ & \xi_d \geq 0, \end{aligned} \tag{3}$$

where ξ_d is a slack variable for soft classification which handles misclassified data samples or anomalies in the data set, and C is a parameter that determines the importance of the outliers. Furthermore, the radial basis function is chosen as a kernel $\phi(x_d)^{\mathrm{T}} \phi(x_d')$ [13].

Hybrid Classification

The hybrid classification extends the threshold-based signal monitoring with machine learning, such that the amount of false negative classified labels is minimized and less manual adjustments by experts are necessary. The workflow of the hybrid classification approach is shown in Fig. 2. For systems with a high repetition rate it is important to make threshold adjustments automatically. In the hybrid classification approach this adjustment is handled by an SVM.

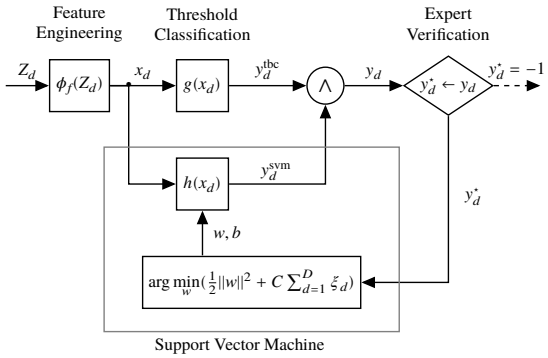


Figure 2: Hybrid classification approach. Similarly to the threshold based approach, first the features are calculated, then a label is assigned to the signal, which is consequently checked by experts. The SVM learns from the past decisions of the expert, by optimizing the parameter w in the separation hyperplane $h(x_d)$.

Specifically, an SVM performs the continuous threshold adjustment, while the threshold-based signal monitoring application operates with fixed thresholds from experts. In the initial phase, the available historical data is used to determine the parameters w and b of the SVM separation hyperplane $h(x_d)$. For each new batch d, the threshold-based classification provides y_d^{tbc} and the SVM classification determines y_d^{svm} from x_d. The output y_d is then determined by combining both outputs with a logical AND, i.e,

$$y_d = \begin{cases} 1 & \text{if } y_d^{\mathrm{tbc}} = 1 \wedge y_d^{\mathrm{svm}} = 1 \\ -1 & \text{otherwise,} \end{cases} \tag{4}$$

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB345

which is verified by experts. Once the experts have evaluated the condition of the component, the corresponding label is added to the training set and the SVM parameters w and b are recalculated. Accordingly, the new label is used as feedback for future decisions of the SVM.

APPLICATION OF THE MODEL

In this section the previously applied hybrid model is applied to the classification of QH breakdowns in the main dipoles of the LHC. The QH discharges are currently analyzed by the Quench Heater Discharge Analysis (QHDA) tool, which groups QH discharges into "healthy" and "faulty" with a threshold-based classification system. An extensive analysis takes place following each quench event before the main dipole can be powered again. In case of a quench in one of the main dipoles, the experts have several hours to check a "faulty" classification of a QH discharge before the magnets can be powered again, which is why a false negative classification (damage predicted while no damage) has a limited impact on the availability of the LHC. A false positive classification (no damage predicted while damage) has to be avoided by all means.

The QHDA tool validates the QH discharges using the following features [2]:

1. *Steady state voltage level:* The initial and final values of the voltage are used.
2. *Characteristic time of the pseudo-exponential decay:* The characteristic time of the pseudo-exponential decay is determined from the voltage and the current signals during the QH discharge.
3. *Steady state resistance level:* The initial resistance of the QH strip is determined from the voltage and current signals.
4. *Signal comparison:* The signals and the above features are compared sample-wise to the reference discharge of the corresponding QH circuit.

Some failures and precursors of failures in the QH circuits are difficult to detect with a threshold based method because they might correlate with other characteristics, which are not verified or which are sensitive to case by case variations. For example the initial resistance of the QH strip is calculated from the voltage and current signals at the start of the QH discharge. Due to the properties of the QH circuit there can be a switch-on delay, oscillations, and noise in both signals, which can cause variation in the calculated initial resistance from discharge to discharge.

Therefore, it was studied, whether a hybrid classification could identify such correlations and if it can consequently decrease the amount of false positive classifications.

RESULTS

The QHDA is implemented into an environment called *LABView*. As machine learning algorithms are commonly implemented in *python*, the features of the QHDA tool were reimplemented in the environment of the "LHC Signal Monitoring Project" [14] to recreate the discussed approaches. The values of the fixed thresholds have been set by experts and the hyperparameters of the SVM have been optimized using training data. The hybrid approach is implemented as stated before, i.e. combining the threshold-based (TB) classification with the SVM classification. The data-set contains stored discharges from 2014 to 2018. 3130 QH discharges have been labeled as "healthy" and come from 1230 main dipole magnets. 116 discharges were labeled as "faulty" and come from 68 different dipole magnets. This data-set was labeled by experts, who classified each discharge, marking even small deviations as "faulty".

Table 1 compares the different methods by their performance. The true positive (TP) rate is the fraction of correctly identified "healthy" discharges relative to the total amount of "healthy" discharges. On the other hand the false positive (FP) rate defines the amount of falsely labeled "healthy" discharges relative to all "faulty" labeled discharges. The TB model indicates the rebuilt QHDA tool in python.

Table 1: Results of Different Performance Measures

Method	TP rate	FP rate
QHDA model	0.999	0.147
TB model (rebuilt QHDA)	0.993	0.078
SVM model	0.991	0.131
Hybrid model	0.989	0.024

From Table 1 it can be seen, that the TB model differs from the QHDA, due to slight variations in the calculation methods, but they both have a relatively good TP rate and a relatively bad FP rate. The hybrid model demonstrates a significantly improved FP rate, while it shows a small degradation in the TP rate. The FP rate of 14.7% for QHDA does not mean that 14.7% of the recorded discharges caused damage to the QHs or the magnets, but indicates the fraction of cases, which required further investigations by an expert. Consequently, experts were only missing 2.4% of the cases, which the hybrid model classified as "faulty". This shows that machine learning improves the identification of cases, which experts need to investigate compared to simpler TB algorithms. Furthermore, ML with an SVM will remember previous expert decisions.

CONCLUSION

In this paper a promising concept for complementing traditional threshold-based LHC signal monitoring tools with machine learning is presented, illustrated by the example of QH signal analysis. This is achieved by building an ensemble of the existing signal monitoring application and an SVM, which is trained on historical data. The conducted analysis showed that this hybrid classification approach reduced the FP rate of the existing QHDA tool from 14.7% to 2.4%. Furthermore, the new approach allows the automatic incorporation of expert decisions into the classification process.

Overall, it has been demonstrated that even a limited amount of historical data can be beneficial for signal monitoring applications through the support of machine learning.

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB345

REFERENCES

[1] R. Schmidt *et al.*, "LHC machine protection," in *Proc. 22nd Particle Accelerator Conf. (PAC'07)*, Albuquerque, NM, USA, Jun. 2007, paper TUZAC03, pp. 878–882.

[2] Z. Charifoulline *et al.*, "Overview of the Performance of Quench Heaters for High-Current LHC Superconducting Magnets," *IEEE Transactions on Applied Superconductivity*, vol. 27, no. 4, pp. 1–5, 2017, doi:10.1109/TASC.2016.2642991

[3] Z. Charifoulline *et al.*, "Resistance of Splices in the LHC Main Superconducting Magnet Circuits at 1.9 K," *IEEE Transactions on Applied Superconductivity*, vol. 28, no. 3, pp. 1–5, 2018, doi:10.1109/TASC.2017.2784355

[4] E. Holzer *et al.*, "Beam Loss Monitoring System for the LHC," in *Proc. IEEE Nuclear Science Symposium Conference Record*, Fajardo, PR, USA, 2005, pp. 1052–1056, doi:10.1109/NSSMIC.2005.1596433

[5] J. Wenninger, "Machine Protection and Operation for LHC," *Proceedings of the Joint International Accelerator School: Beam Loss and Accelerator Protection*, vol. 2, no. 1, pp. 377–401, 2016, doi:10.5170/CERN-2016-002.377

[6] U. Gentile and L. Serio, "A Machine-Learning Based Methodology for Performance Analysis in Particles Accelerator Facilities," in *2017 European Conference on Electrical Engineering and Computer Science (EECS)*, Bern, Switzerland, 2017, pp. 90–95, doi:10.1109/EECS.2017.26

[7] M. Wielgosz, A. Skoczeń, and M. Mertik, "Using LSTM recurrent neural networks for monitoring the LHC superconducting magnets," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 867, no. 1, pp. 40–50, 2017, doi:10.1016/j.nima.2017.06.020

[8] A. Kharrat, K. Gasmi, M. B. Messaoud, and M. Abid, "A Hybrid Approach for Automatic Classification of Brain MRI Using Genetic Algorithm and Support Vector Machine," *Leonardo journal of sciences*, vol. 17, no. 1, pp. 71–82, 2010.

[9] M. Seera, C. P. Lim, S. Nahavandi, and C. K. Loo, "Condition monitoring of induction motors: A review and an application of an ensemble of hybrid intelligent models," *Expert Systems with Applications*, vol. 41, no. 10, pp. 4891–4903, 2014, doi:10.1016/j.eswa.2014.02.028

[10] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010, doi:10.1007/s10462-009-9124-7

[11] J. Surowiecki, *The wisdom of crowds*, 1. ed. New York, NY: Anchor Books, 2005.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi:10.1145/2939672.2939785

[13] C. M. Bishop, *Pattern recognition and machine learning*. New York, NY, USA: Springer, 2006.

[14] M. Maciejewski, Z. Charifoulline, A. Verweij, and P. Hagen, Signal Monitoring, https://sigmon.web.cern.ch/about.

Paper 2

Machine Learning Models for Breakdown Prediction in RF Cavities for Accelerators

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB344

MACHINE LEARNING MODELS FOR BREAKDOWN PREDICTION IN RF CAVITIES FOR ACCELERATORS

C. Obermair*[1], A. Apollonio, T. Cartier-Michaud, W. L. Millar, F. Pernkopf[1], W. Wuensch, N. Catalan-Lasheras, L. Felsberger
CERN, Geneva, Switzerland
[1]also at Graz University of Technology, Graz, Austria

Abstract

Radio Frequency (RF) breakdowns are one of the most prevalent limits in RF cavities for particle accelerators. During a breakdown, field enhancement associated with small deformations on the cavity surface results in electrical arcs. Such arcs degrade a passing beam and if they occur frequently, they can cause irreparable damage to the RF cavity surface. In this paper, we propose a machine learning approach to predict the occurrence of breakdowns in CERN's Compact LInear Collider (CLIC) accelerating structures. We discuss state-of-the-art algorithms for data exploration with unsupervised machine learning, breakdown prediction with supervised machine learning, and result validation with Explainable-Artificial Intelligence (Explainable AI). By interpreting the model parameters of various approaches, we go further in addressing opportunities to elucidate the physics of a breakdown and improve accelerator reliability and operation.

INTRODUCTION

The novel RF cavities of CERN's Compact LInear Collider (CLIC) are designed for high gradient operation at ~100 MV/m [1]. Even though RF cavities are operated in vacuum, local field emissions can cause arcs and breakdowns of the electric field in the cavity which have a negative effect on the cavity surface material. The frequency of these arcs, described by the breakdown rate, is the main limitation to increase the electric field in an RF cavity during conditioning and operation. While historically RF structures have been conditioned in a manual way by machine operators, an automated conditioning algorithm is in place at the CLIC test stand to gradually increase the field gradient while maintaining a pre-defined target breakdown rate [2]. These conditioning efforts set the limit to the gradient due to field emission, caused by the geometrical defects of the surfaces and the RF power flow, to reduce the likelihood of a breakdown [3, 4]. Given the limited understanding of the origin and evolution of RF breakdowns, current optimization algorithms aim for a progressive recovery of operating conditions by a temporary limitation of the RF power after a breakdown, but do not avoid breakdowns in the first place. Recently, data-driven machine learning algorithms have been deployed successfully for incorporating sequential dynamics [5, 6] using the large amount of experimental data available. Ongoing efforts already try to predict breakdowns in the RF power source output of CERN's LINAC 4 [7], or to classify superconducting RF faults at Jefferson Laboratory [8].

This paper gives an overview of several data-driven methods for RF breakdown analysis, specifically suited to the properties of the measurement data of the CLIC XBOX-2 test stand at CERN. The paper provides an introduction, comparison, and hands-on experience of existing data-driven modeling approaches to non-machine learning experts. It provides RF physicists and engineers with machine learning based tools, which allow to gain insights in observing abnormal behaviours. Finally, first results of these methods applied to the CLIC XBOX-2 test stand data are presented.

The paper is structured as follows. First, the properties of the CLIC XBOX-2 test stand and its historical data are described. Consecutively, a broad overview of existing machine learning algorithms, suitable for breakdown prediction in the test stand, is given. Finally, their strengths and the limitations are discussed, first results are presented and an outlook is given.

TEST STAND SETUP

The CERN XBOX-2 test stand is part of the CLIC e+e- collider research program for high gradient acceleration in high gradient structures. It is one of three high power test stands at CERN and its primary objective is to study the RF breakdown phenomenon. A low level radio frequency generator creates a 1.5 µs long, 12 GHz phase-modulated pulse. This pulse is amplified by a klystron and a pulse compressor and is then transferred through a copper wave guide to the RF cavity [9]. A diagram of the high-power portion of the test stand layout is provided in Fig. 1. The RF cavity is represented as Device Under Test (DUT). The signals from the upstream and downstream Faraday cups, which measure the dark current in the structure, are symbolized by the blue arrows labelled DC UP and DC DOWN.

METHODOLOGIES

In order to give a hands-on overview, in this paper, the choice of an algorithm is governed by the chronological order of the data processing, i.e. transformation, exploration, modeling, and explanation of RF cavity specific data.

Existing open-source libraries are used instead of handcrafted methods in all processing steps, because the engineering and the maintenance of customized methods is time consuming. For the labeled measurement data from the XBOX-2 test stand, dedicated toolboxes are used for feature

* christoph.obermair@cern.ch

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB344

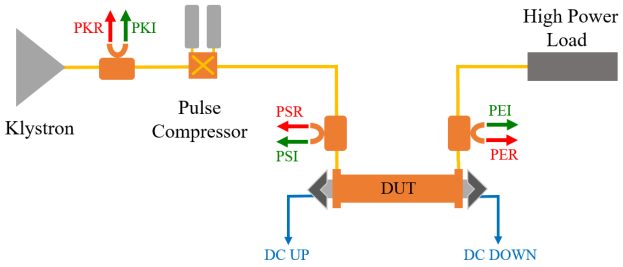


Figure 1: Schematic layout of CERN's Xbox-2 test stand. The red and green arrows show the reflected and forward RF signals, respectively, which are sampled via directional couplers.

calculation [10], time series classification [11], and interpretation of model predictions [12].

Transformation

The sensor data from the XBOX-2 high power tests is divided into so-called *trend* data and *event* data. While the trend data contains single scalar features (e.g. temperatures), the event data contains time series signals, generally sampled with a frequency of 1.6 GHz. In total there are 90 GB of data available from a period of six months of XBOX-2 operation in 2018. These data do not only contain runs in which the operational setting was stable, but also commissioning data with variable operational settings. Thus, it is crucial to initially clean the data, create fast queries, memory efficient storage and file formats with diverse usability. Figure 2 shows the condition summary of the data, where the runs with stable operational settings are highlighted in yellow and the cumulative number of breakdowns is shown in red. The plot further shows the input power in blue, and the pulse width of the input signal in green.

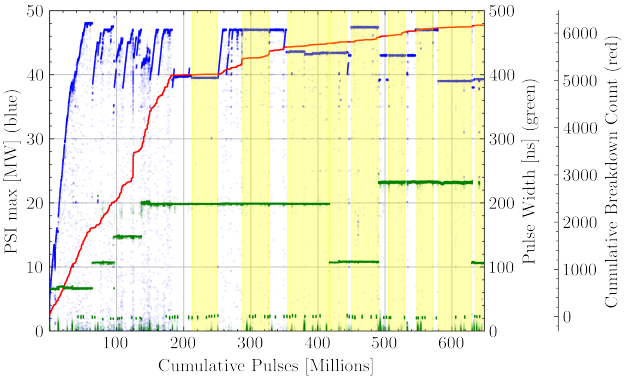


Figure 2: Condition summary of available data. The yellow area represents the runs during which the operational settings were kept stable.

A breakdown results in a burst of current in the cavity, which can be detected by the Faraday cups next to the structure. Therefore, for each event data signal, a label *healthy* ($y = 1$) and *breakdown* ($y = 0$) is assigned by the XBOX-2 experts by setting a threshold on the DC (Faraday cup) signals and the reflected signals. However, as the DC signals are the most reliable filter for structure breakdowns, the RF signals accounting for the reflected power in the structures are not considered for breakdown prediction. Specifically, this means a signal is considered a breakdown, if one of the DC time series signals goes below -0.05 A. In addition, a label is considered a so-called *follow-up breakdown*, if there has already been a breakdown within less than a minute from its occurrence. After filtering out the test stand commissioning data, where most of the breakdowns occurred, 124,448 healthy events and 479 breakdown events, out of which 250 are follow-up breakdowns, remained for further analysis. This class imbalance is tackled by only taking a sub-set of healthy signals and by assigning class weights to the breakdown events during optimization of the algorithm and during computation of the performance measure.

Merging and synchronizing the trend data with the event data is a critical data transformation step. Due to its high sampling frequency, an event data signal with up to 3200 sample points is stored every minute. Exceptions are breakdown events, where the prior two event data signals are stored each time a pulse is injected into the RF cavity. The scalar values of the trend data take up much less space, and are therefore stored every second. During merging of event data and trend data, causality is ensured by always taking the closest information in the past, not in the future.

Exploration

During the exploration phase the goal is to get a quick initial understanding of the data and to validate the transformation step, i.e. if the preceding data cleaning was successful. If there are still outlier signals, which are fundamentally different from the other signals, they have to be understood and, if applicable, neglected. Ideally, a 2D-representation should be found for each event in the high dimensional data, without losing any information due to the dimension reduction. This allows to see correlations and clusters within the representations in one glance. Several unsupervised machine learning methods aim to determine low-dimensional representations from the high dimensional data, including but not limited to principle component analysis [13], stochastic neighbor embeddings [14], and representation learning methods based on neural networks [14–16].

An example is shown in Fig. 3, where the XBOX-2 trend data is transformed into a two dimensional space with 2D-tSNE [14]. 2D-tSNE transforms pairs of data points to joint probabilities, where close points have high probabilities and points which are far apart have low probabilities. Consequently, the Kullback-Leibler divergence within the joint probabilities of the low-dimensional representations and the high-dimensional data is minimized iteratively. While the axes lose their physical meaning during the dimension reduction, one can clearly see clusters of breakdowns and healthy signals in the left plot and the nine different stable runs in the right plot of Fig. 3. Neither the information of the label, nor the information of the runs were given to the algorithm during training.

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB344

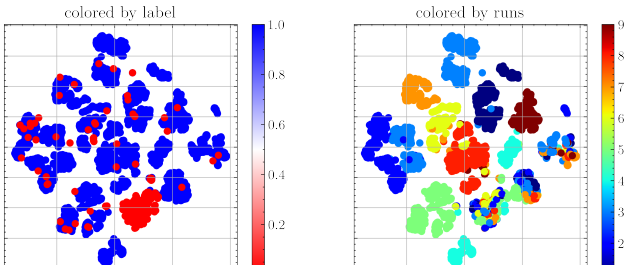


Figure 3: 2D-tSNE of XBOX-2 trend data during stable operation. The algorithm was able to distinguish between healthy and breakdown signals (left) and between stable runs (right). No information about the labels was given to the algorithm.

Modeling

We propose two supervised machine learning stages. First, the behaviour of the trend data over time is investigated. This means that a window, covering a certain time-span of data, is moved over the data-set. For each window a prediction is made if a breakdown will occur in a certain time period. In this step, no shuffling of time series data is allowed due to the sequential dependency. Recurrent neural networks, like long short-term memory networks [17], are especially suited to process this temporally dynamic behavior due to their recurrent neuron connections. The model is trained with several rounds of leave-one-out cross-validation. One round of cross-validation involves splitting the data-set into a training, validation and test set, based on the given runs. This process is repeated until each run was used as training, validation and test set. In a second stage, it is assumed that only the signals before a breakdown are essential to predict the breakdown. Therefore, the signals from the event data are taken and treated independently with their own label *breakdown in the next pulse*. This has the advantage that convolutional neural networks can be used to classify the time-series signals [18]. Additionally, the data-set is shuffled, and the signals are randomly split into training, validation, and test sets.

Explanation

To increase the reliability of a system, understanding why the prediction was made, i.e. looking for a precursor, is often more important than the prediction itself. Especially when designing upgrades of existing systems, a deep understanding of the root cause of the failures can be an invaluable asset. As data-driven models are often black-boxes, explainable-AI does not only help the user to better interpret the behaviours of the models, but it also helps to build trust in the prediction, to validate the results, and to find possible errors within the earlier data processing steps. One can either explore each prediction separately to gain trust in a prediction (instance wise explanation) [19–22], or investigate all predictions to gain trust in a model (population wise explanation) [23]. Both approaches are applicable for explaining predictions of RF cavity breakdowns.

RESULTS & CONCLUSION

Table 1 shows the results of the trained supervised models. The balanced accuracy is used for taking into account the strong class imbalance. It is calculated by averaging the fraction of correctly categorization breakdowns and healthy signals.

Table 1: Balanced Accuracy of Classifying and Predicting Breakdowns With XBOX-2 Data. The Separation Indicates Different Results on Breakdowns / Follow-up Breakdowns

	Classification of Breakdowns	**Prediction of Breakdowns**
Trend Data	100%	91%
Event Data	100%	65% / 98%

The classification step is required for the validation of the algorithms applied on the trend and event data. The balanced accuracy of 100% for trend and event data in the classification shows the successful validation of the algorithms.

The models achieved a balanced accuracy of 91% for predicting breakdowns in the next pulse using trend data. Here, explainable-AI showed that the models made decisions mainly by using the vacuum signals. After further investigation, it was found that a rise in the vacuum pressure mostly occurred just before a breakdown and not only after a breakdown, as generally assumed. This rise in vacuum pressure might be due to small breakdowns happening just before a major breakdown. Further experiments in the test stand are ongoing to validate this result and exclude any artefacts due to signal timing in the experimental setup.

By using the time-series signals of the event data, a balanced accuracy of 65% was achieved for predicting breakdowns, and 98% for predicting follow-up breakdowns. Here, explainable-AI indicates precursors in multiple ways, pointing to the most important part of each measurement, or indicating the three most similar events present in the rest of the data set.

Using this method, an additional precursor has been identified. Faraday cup signals with a small spike, which occurs relatively late in the signal but does not reach the breakdown threshold, often leads to consecutive breakdowns in the next pulse. Following further validations of these results, an operational tool for breakdown reduction based on the described machine learning methods will be developed.

REFERENCES

[1] E. Sicking and R. Ström, “From precision physics to the energy frontier with the Compact Linear Collider”, *Nature Physics*, vol. 16, pp. 386–392, 2020. doi:10.1038/s41567-020-0834-8

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-MOPAB344

[2] W. Wuensch *et al.*, "Experience Operating an X-band High-Power Test Stand at CERN", in *Proc. 5th Int. Particle Accelerator Conf. (IPAC'14)*, Dresden, Germany, Jun. 2014, paper WEPME016, pp. 2288–2290.

[3] A. Grudiev and W. Wuensch, "A New Local Field Quantity Describing the High Gradient Limit of Accelerating Structures", in *Proc. 24th Linear Accelerator Conf. (LINAC'08)*, Victoria, Canada, Sep.-Oct. 2008, paper THP063, pp. 936–938.

[4] A. Descoeudres, Y. Levinsen, S. Calatroni, M. Taborelli, and W. Wuensch, "Investigation of the dc vacuum breakdown mechanism", *Physical Review Special Topics - Accelerators and Beams*, vol. 12, p. 092001, 2009. doi:10.1103/PhysRevSTAB.12.092001

[5] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", in *Proc. of the Annual Conf. of the Int. Speech Communication Association(PACISCA'14)*, USA, 2014, pp. 338-342.

[6] N. Catalan-Lasheras *et al.*, "Commissioning of XBox-3: A Very High Capacity X-band Test Stand", in *Proc. 28th Linear Accelerator Conf. (LINAC'16)*, East Lansing, MI, USA, Sep. 2016, paper TUPLR047, pp. 568–571.

[7] Y. Donon *et al.*, "Extended anomaly detection and breakdown prediction in LINAC 4's RF power source output", in *Proc. of Int. Conf. on Information Technology and Nanotechnology (ITNT'20)*, Samara, Russia, May 2020, pp. 1-7. doi:10.1109/ITNT49337.2020.9253296

[8] C. Tennant, A. Carpenter, T. Powers, A. S. Solopova, L. Vidyaratne, and K. Iftekharuddin, "Superconducting radio-frequency cavity fault classification using machine learning at Jefferson Laboratory", *Physical Review Accelerators and Beams*, vol. 23, p. 114601, 2020. doi:10.1103/PhysRevAccelBeams.23.114601

[9] B. J. Woolley, "High Power X-band RF Test Stand Development and High Power Testing of the CLIC Crab Cavity", Ph.D. dissertation, Lancaster University, UK, 2015.

[10] Tsfresh, https://tsfresh.readthedocs.io

[11] Github, https://github.com/hfawaz/dl-4-tsc

[12] Shap, http://shap.readthedocs.io

[13] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37-52, 1987. doi:10.1016/0169-7439(87)80084-9

[14] G. Hinton and R. Sam, "Stochastic Neighbor Embedding", in *Proc. of the 15th Int. Conf. on Neural Information Processing Systems (NIPS'02)*, Cambridge, MA, USA, Jan. 2002, pp. 857–864.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013. doi:10.1109/TPAMI.2013.50

[16] J.-Y. Franceschi, A. Dieuleveut, and M. Jaggi, "Unsupervised Scalable Representation Learning for Multivariate Time Series", in *Proc. 33rd Conf. on Neural Information Processing Systems (NeurIPS'19)*, Vancouver, Canada, Dec. 2019, pp. 4650–4661.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, vol. 9, pp. 1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735

[18] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification", *Data Mining and Knowledge Discovery*, vol. 33, pp. 917–963, 2019. doi:10.1007/s10618-019-00619-1

[19] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, "How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods", in *34th Conf. on Neural Information Processing Systems (NeurIPS'20)*, Vancouver, Canada, Dec. 2020, pp. 4211–4222.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Proc. 31st Conf. on Neural Information Processing Systems (NIPS'17)*, Long Beach, CA, USA, Dec. 2017.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", 2016. arXiv:1602.04938

[22] S. Tonekaboni, S. Joshi, K. Campbell, D. K. Duvenaud, and A. Goldenberg, "What went wrong and when? Instance-wise feature importance for time-series black-box models", in *Proc. 34th Conf. on Neural Information Processing Systems (NeurIPS'20)*, Vancouver, Canada, Dec. 2020, pp. 262–269.

[23] M. A. Wojtas and K. Chen, "Feature Importance Ranking for Deep Learning", in *Proc. 34th Conf. on Neural Information Processing Systems (NeurIPS'20)*, Vancouver, Canada, Dec. 2020, pp. 5105–5114.

Paper 3

Data Augmentation for Breakdown Prediction in CLIC RF Cavities

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOMS054

DATA AUGMENTATION FOR BREAKDOWN PREDICTION IN CLIC RF CAVITIES

H. S. Bovbjerg *[1,2], C. Obermair[1,3], A. Apollonio[1], T. Cartier-Michaud[1], W. Millar[1], Z.H. Tan[2], M. Shen[2], D. Wollmann[1]
[1]CERN, Geneva, Switzerland [2]Aalborg University, Aalborg, Denmark [3]TU Graz, Graz, Austria

Abstract

One of the primary limitations on the achievable accelerating gradient in normal-conducting accelerator cavities is the occurrence of vacuum arcs, also known as RF breakdowns. A recent study on experimental data from the CLIC XBOX2 test stand at CERN proposes the use of supervised machine learning methods for predicting RF breakdowns. As RF breakdowns occur relatively infrequently during operation, the majority of the data was instead comprised of non-breakdown pulses. This phenomenon is known in the field of machine learning as class imbalance and is problematic for the training of the models. This paper proposes the use of data augmentation methods to generate synthetic data to counteract this problem. Different data augmentation methods like random transformations and pattern mixing are applied to the experimental data from the XBOX2 test stand, and their efficiency is compared.

INTRODUCTION

The RF cavities of the Compact LInear Collider (CLIC) are designed to operate at a gradient of ~100 MV/m [1]. One of the primary limitations on the achievable gradient in normal conducting RF cavities is the occurrence of RF breakdowns, which can degrade a passing beam and potentially result in damage to the cavity surface [2–4]. In order to minimize the impact of breakdowns during the cavity commissioning and operation, CERN's CLIC test stands [5] employ an automatic conditioning algorithm [6, 7]. The algorithm monitors how frequently breakdowns occur during operation and dynamically adjusts the gradient based on a preset breakdown-rate threshold [8]. In this approach, the handling of breakdowns is therefore purely reactive, thus breakdowns cannot be prevented beforehand.

In a recent study, a deep learning approach was proposed with the goal of (1) performing data-driven breakdown investigation and (2) studying the possibility of adopting a predictive conditioning algorithm. The study was based on historical data of the CERN XBOX2 test stand, consisting of 124 505 healthy RF pulses and 479 breakdown events [9].

Previously, it has been noted that breakdowns occur predominantly in groups as opposed to isolated, single events. This observation has led to the classification of breakdown events as either *primary breakdowns*, which are purely stochastic, and *followup breakdowns*, which are thought to be a consequence of the previous breakdown [10]. Using the XBOX2 data, neural networks were able to predict the occurrence of followup breakdowns. However, the prediction accuracy varied depending on different data used for the prediction, e.g. for different adopted parameters for cavity powering. This variation indicates that the models were not able to generalize well to unseen data [11]. Specifically, the bad generalization is due to the low number of breakdown events compared to the number of healthy events, i.e. the so-called high *class imbalance*. We therefore investigated the use of time series data augmentation methods for improving the generalization capabilities of CLIC breakdown prediction. The basic principle of these methods involves generating synthetic patterns that resemble real data to better represent the underlying distribution of the underrepresented class in the data set. This is an established practice for image recognition tasks [11–13] and is also used for speech and audio [14, 15].

The paper is structured as follows: first, a summary of the prior work is given, including a description of the data and model used in our study. Next, an overview of the augmentation methods used in this paper is presented. Finally, the conducted experiments are described, and their results are discussed.

PRIOR WORK

This section summarizes the prior work which this work builds upon, including a description of the data set used in the study, and a description of the RF breakdown prediction models used.

XBOX2 Data Set

The XBOX2 test stand is one of three experiments used to test the prototype 12GHz RF components for the CLIC project at CERN. Fundamentally, the test stand is composed of a 50 MW klystron, pulse compressor, and high-power RF load. A more detailed description of the setup is available elsewhere [7, 9].

In 2018 this test stand produced 90 GB of data during an operational period of six months, consisting of so-called trend and event data [9]. The trend data contains 30 different scalar values such as temperatures and pressures measured at different locations in the test-stand. The event data consists of time-series measurements of the RF signals at different locations in the waveguide network and the current detected by two Faraday cups. A summary of the data is given in Fig. 1. Here, two features of the forward travelling wave signal F2 (see fig. 2), namely the maximum (blue) and the pulse width (green), are shown with respect to the RF cavity pulses. Additionally, the cumulative breakdown count (red)

* holger.severin.bovbjerg@cern.ch

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOMS054

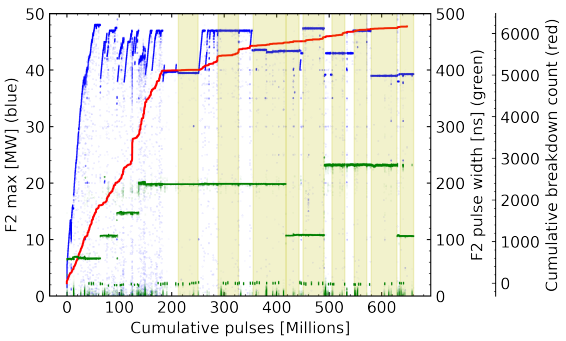


Figure 1: Overview of the conditioning period, of all data analyzed [9]. It shows the maximum of the power amplitude of the forward travelling wave signal F2 (blue), its pulse width (green), and the cumulative breakdown count (red).

is plotted. The yellow area represents the periods with constant operational settings used for further analysis, leading to a total of 479 breakdowns and 124 505 healthy RF pulses, as not every pulse is stored. Given the previously observed probabilistic behavior of breakdowns, the data is further divided into 229 primary and 250 followup breakdowns. Primary breakdowns were defined as not having occurred within 3000 pulses of the previous breakdown, corresponding to one minute of operation in the test stand, which has a pulse repetition rate of 50 Hz.

Modelling of RF Breakdowns

In [9], a number of neural network architectures were investigated to predict breakdowns using trend data and event data. These two experiments were further split into the prediction of primary breakdowns and followup breakdowns. Formally, the prediction of breakdowns is defined as finding a model $f(\cdot)$ that uses the observed data $\mathbf{x}_i$ to predict the label (healthy or breakdown) of the next time stamp y_{i+1}, where i is the current time step.

The model performance is measured using the Area under the Receiver operating characteristics curve (AR) [16]. This score is defined as the probability that a model will classify a randomly selected breakdown event as more likely to be a breakdown than a randomly selected healthy event. An AR score of 100% means that the model is able to perfectly predict the class labels, and a score of 0% corresponds to a classifier which predicts all labels wrong.

Primary breakdowns proved to be difficult to predict with available event data, whereas it was possible to predict followup breakdowns with an AR score of up to 89.7% ± 8.1%. We aim at further improving the Fully Convolutional Network (FCN), achieving this result, with data augmentation.

DATA AUGMENTATION

The XBOX2 data consist of a number of time series, therefore we focus on time series augmentation methods, which can generally be divided into four categories: random transformations, pattern mixing, generative models and decomposition models [11]. In this study, we only consider random transformation methods and pattern mixing. We do not consider generative models due to the computational cost and their high number of parameters. Furthermore, due to the non-periodic nature of the XBOX2 data, decomposition models are deemed inapplicable. Illustrations of all applied methods are seen in Figure 2.

Random Transformation

Random transformation methods apply different types of transformations to the data, in order to generate new synthetic samples. Random transformation methods assume that the transformations are representative of the data characteristics [11], i.e. they can be introduced without changing the fundamental nature of the signals. Typically, augmentation methods alter the values, the time steps or the frequencies in a signal, i.e. transformations take place in the magnitude, time, or frequency domain. In the case of the XBOX2 data, frequency transformations are not applicable, as the data is not periodic.

A simple random transformation method in the magnitude domain is *noise addition*, also known as *jittering*. Here, a noise vector $\boldsymbol{\alpha}$ is sampled from a zero mean Gaussian $\sim \mathcal{N}(0, \sigma^2)$, which is then added to a data sample $\mathbf{x}$ to generate a synthetic sample $\mathbf{x}'$ such that $\mathbf{x}' = \mathbf{x} + \boldsymbol{\alpha}$. Adding noise has been shown to improve generalization of neural networks [17].

Another similar strategy, known as *magnitude scaling* [18], scales the data sample by a Gaussian scaling vector $\boldsymbol{\beta} \sim \mathcal{N}(1, \sigma^2)$, such that $\mathbf{x}' = \mathbf{x} \cdot \boldsymbol{\beta}$. A more advanced version of this approach is known as *magnitude warping* [18]. Here the scaling vector is based on interpolation from a cubic spline S with k knots, with the knots being drawn from a Gaussian $\sim \mathcal{N}(1, \sigma^2)$.

Random transformation methods that act in the time domain include warping and slicing methods. *Window slicing* generates new samples by only selecting a certain percentage W of the available samples, and interpolating back to the original number of samples. *Warping* in time involves perturbing the individual data point of a sample in time. Given a warping function τ, defined by a cubic spline S with k knots drawn from a Gaussian distribution $\sim \mathcal{N}(1, \sigma^2)$, a new sample is found as $\mathbf{x}' = x_{\tau(1)}, \ldots, x_{\tau(t)}, \ldots, x_{\tau(T)}$, with T being the sample length.

Pattern Mixing

Pattern mixing techniques seek to generate synthetic samples by mixing features of multiple data samples. In its simplest form, pattern mixing takes the mean between two or more signals of the same class. However, this method might remove distinguishing features from the signal, due to smoothening from the mean operator.

A popular method for pattern mixing is known as Synthetic Minority Oversampling Technique (*SMOTE*) [19]. The SMOTE method takes a sample of the minority class $\mathbf{x}$ and randomly selects a k-nearest neighbor $\mathbf{x}_{\mathbf{NN}}$. The absolute difference between them is then found and scaled by a random scalar $\lambda \sim \mathcal{U}(0, 1)$, and the new sample is found as $\mathbf{x}' = \mathbf{x} + \lambda|\mathbf{x} - \mathbf{x}_{\mathbf{NN}}|$.

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOMS054

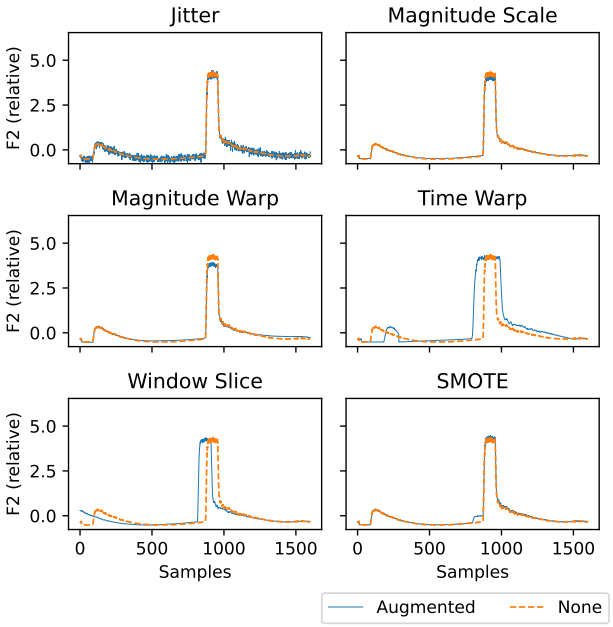


Figure 2: Illustration of augmentation techniques applied to a forward travelling wave signal F2 from the XBOX2 data set.

EXPERIMENTS

To test whether data augmentation is beneficial for the prediction of RF breakdowns, a series of simulation experiments have been carried out. For each of the selected data augmentation methods, we train the FCN model following the approach of [9]. Each augmentation method includes a number of hyperparameters, which we choose based on recommended values from literature [11].

In all our data augmentation methods, we oversample the minority class and take only 2.5% of healthy events, i.e. 3113 events, similarly to prior work [9]. Considering the whole data set, we augment 3113 healthy and all 250 followup breakdowns, to acquire 3113 healthy and 3113 followup breakdowns. Data augmentation aims to remove the class imbalance, making class weighting used in previous work [9], not always necessary. The best results of each method are summarized in Table 1. Methods with class weighting are marked with (*).

To fairly assess the model performance with data augmentation, a *train on synthetic test on real* paradigm is used. This means that the models are trained on a training set containing synthetic data, however, the validation and test set is kept untouched. In this manner, the performance using data augmentation can directly be compared to the baseline model trained without data augmentation.

The periods of stable operation are used for k-fold cross-validation. This means that one group is set aside as a validation set, using the rest for training. Each stable operation is used as a validation set once. The mean AR score is then reported as AR_μ with standard deviation AR_σ. After fine-tuning manual model parameters, the model is finally trained on both training and validation set, and tested on an unseen stable operation period with a performance AR_t.

Table 1: Best AR Scores for Various Implemented Augmentation Methods

Augmentation	AR_μ [%]	AR_σ [%]	AR_t [%]
None*	89.7	8.1	91.1
Jitter	90.0	2.9	84.2
Magnitude Scaling	89.4	5.1	87.6
Magnitude Warp	90.4	4.9	86.0
Time Warp	90.8	4.6	84.6
Window Slicing*	89.4	5.0	90.4
SMOTE*	91.0	5.2	89.8

RESULTS & DISCUSSION

In Table 1, we present the results obtained from applying data augmentation methods to the XBOX2 data when predicting followup breakdowns. When comparing the results using no data augmentation to the results with data augmentation, a slight increase in the mean AR score on the validation set is seen for jittering, magnitude warping, time warping and SMOTE. Magnitude scaling and window slicing instead show a slight decrease. The SMOTE method achieves the largest improvement over no data augmentation and yields an improvement of 1.3%, when keeping the class weighting from the previous study. The best result achieved with no class weighting was for time warping, with an improvement of 1.1%.

Looking at the standard deviation, all augmentation methods yield a significant decrease. This means that the performance of the trained model varies less on different validation sets when using data augmentation, and that the models are able to generalizes better, independently of the stable operation period. The best performance, with respect to AR_σ, was achieved by jittering which decreased the standard deviation by 5.2% compared to no augmentation, yielding a standard deviation of 2.9%. Note, that AR_t is only used to validate the model's generalization capabilities by testing whether AR_t is within $AR_\mu \pm 2AR_\sigma$.

CONCLUSION

In this paper, we investigated different techniques to improve existing RF breakdown prediction models through the use of data augmentation methods applied to time series data from CERN's XBOX2 test stand. We conclude that data augmentation improves the standard deviation of our model independent of the technique, making the used model more robust and generic. The performance of the model, however, only improve slightly dependent on the technique. The best performance was achieved using the SMOTE method, keeping the class weighting from the original study. SMOTE improved the average model performance by 1.3% and decreased the standard deviation by 2.9%. The achieved results provide new insights for the development of a proactive and dynamic conditioning algorithm for CLIC RF cavities.

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOMS054

REFERENCES

[1] CLICdp Collaboration, *The Compact Linear Collider (CLIC) - 2018 Summary Report*, P. Burrows, Ed. 2018, vol. 2, doi:10.23731/CYRM-2018-002

[2] H. Wiedemann, *Particle Acelerator Physics*, 4th ed. Springer, 2015.

[3] B. Woolley *et al.*, "High-gradient behavior of a dipole-mode rf structure," *Physical Review Accelerators and Beams*, vol. 23, 2020, doi:10.1103/PhysRevAccelBeams.23.122002

[4] A. Hassanein *et al.*, "Effects of surface damage on rf cavity operation," *Phys. Rev. ST Accel. Beams*, vol. 9, p. 062 001, 6 2006, doi:10.1103/PhysRevSTAB.9.062001

[5] T. Lucas *et al.*, *High power testing of a prototype clic structure: Td26cc r05 n3*, 2018.

[6] W. Wuensch *et al.*, "Experience Operating an X-band High-Power Test Stand at CERN," in *Proceedings of the 5th Int. Particle Accelerator Conf.*, 2014, doi:10.18429/JACOW-IPAC2014-WEPME016

[7] L. Millar, "Operation of Multiple Accelerating Structures in an X-Band High-Gradient Test Stand," Presented 22 Jul 2021, 2021, https://cds.cern.ch/record/2798232

[8] N. Catalan-Lasheras *et al.*, "Commissioning of XBox-3: A Very High Capacity X-band Test Stand," in *Proc. LINAC'16*, East Lansing, MI, USA, Sep. 2016, pp. 568–571, doi:10.18429/JACoW-LINAC2016-TUPLR047

[9] C. Obermair *et al.*, "Explainable machine learning for breakdown prediction in high gradient rf cavities," 2022, doi:10.48550/ARXIV.2202.05610

[10] W. Wuensch *et al.*, "Statistics of vacuum breakdown in the high-gradient and low-rate regime," *Phys. Rev. Accel. Beams*, vol. 20, p. 011 007, 1 2017, doi:10.1103/PhysRevAccelBeams.20.011007

[11] B. K. Iwana and S. Uchida, "An empirical survey of data augmentation for time series classification with neural networks," *PLOS ONE*, vol. 16, no. 7, e0254841, 2021, doi:10.1371/journal.pone.0254841

[12] Q. Wen *et al.*, "Time series data augmentation for deep learning: A survey," 2021, doi:10.24963/ijcai.2021/631

[13] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014, doi:10.48550/ARXIV.1409.1556

[14] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[15] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. M. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *INTERSPEECH*, 2016.

[16] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi:10.1016/j.patrec.2005.10.010

[17] G. An, "The Effects of Adding Noise During Backpropagation Training on a Generalization Performance," *Neural Computation*, vol. 8, no. 3, pp. 643–674, 1996, doi:10.1162/neco.1996.8.3.643

[18] T. T. Um *et al.*, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," 2017, doi:10.1145/3136755.3136817

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi:10.1613/jair.953

Paper 4

Explainable Machine Learning for Breakdown Prediction in High Gradient RF Cavities

Explainable machine learning for breakdown prediction in high gradient rf cavities

Christoph Obermair[*]
CERN, CH-1211 Geneva, Switzerland and Graz University of Technology, AT-8010 Graz, Austria

Thomas Cartier-Michaud, Andrea Apollonio, William Millar,[†] Lukas Felsberger, Lorenz Fischl,[‡] Holger Severin Bovbjerg,[§] Daniel Wollmann, Walter Wuensch, Nuria Catalan-Lasheras, and Marçà Boronat
CERN, CH-1211 Geneva, Switzerland

Franz Pernkopf
Graz University of Technology, Graz, Austria

Graeme Burt
Cockcroft Institute, Lancaster University, Lancaster, United Kingdom

(Received 10 January 2022; accepted 6 September 2022; published 4 October 2022)

The occurrence of vacuum arcs or radio frequency (rf) breakdowns is one of the most prevalent factors limiting the high-gradient performance of normal conducting rf cavities in particle accelerators. In this paper, we search for the existence of previously unrecognized features related to the incidence of rf breakdowns by applying a machine learning strategy to high-gradient cavity data from CERN's test stand for the Compact Linear Collider (CLIC). By interpreting the parameters of the learned models with explainable artificial intelligence (AI), we reverse-engineer physical properties for deriving fast, reliable, and simple rule–based models. Based on 6 months of historical data and dedicated experiments, our models show fractions of data with a high influence on the occurrence of breakdowns. Specifically, it is shown that the field emitted current following an initial breakdown is closely related to the probability of another breakdown occurring shortly thereafter. Results also indicate that the cavity pressure should be monitored with increased temporal resolution in future experiments, to further explore the vacuum activity associated with breakdowns.

DOI: 10.1103/PhysRevAccelBeams.25.104601

I. INTRODUCTION

In the field of particle accelerators, specially designed metallic chambers known as radio-frequency (rf) cavities are commonly employed to establish electromagnetic fields capable of accelerating traversing particles. The energy gain provided by a cavity is determined by the accelerating gradient, a quantity defined as the longitudinal voltage experienced by a fully relativistic traversing particle normalized to the cavity length. Hence, in linear accelerators (LINACS), any increase in the accelerating gradient translates to a reduced machine length. The continued interest in future colliders and other accelerator applications, where machine size is a key constraint, has continued to drive research in this area. One such example is CERN's Compact LInear Collider (CLIC) project, a proposed future high-energy physics facility that aims to collide positrons and electrons at an energy of 3 TeV. To reach this energy at an acceptable site length and at an affordable cost, the project proposes the use of X-band normal-conducting copper cavities operating at an accelerating gradient of 100 MV/m [1].

One of the primary limits on the achievable accelerating gradient in normal conducting high-gradient cavities is a phenomenon known as vacuum arcing or breakdown [2]. To operate reliably at high accelerating gradients, such cavities must first be subjected to a so-called *conditioning period* in which the input power is increased gradually while monitoring for breakdowns [3–5]. Due to the limited understanding of the origin of rf breakdowns and the inability to predict them, current operational algorithms

[*]christoph.obermair@cern.ch
[†]Also at Cockcroft Institute, Lancaster University, Lancaster, United Kingdom.
[‡]Also at Vienna University of Technology, Vienna, Austria.
[§]Also at Aalborg University, Aalborg, Denmark.

generally act responsively rather than preemptively. Hence, they aim for a progressive recovery of operating conditions by temporarily limiting the rf power following breakdowns [6]. In this paper, we investigate the possibility of employing predictive methods based on machine learning to limit the impact of breakdowns.

Data-driven machine learning algorithms have been successfully deployed in particle accelerator applications for incorporating sequential dynamics using large amounts of available experimental data. Ongoing efforts at CERN have demonstrated the successful use of machine learning for failure analysis in particle accelerators, e.g., to identify and detect anomalies in the rf power source output of LINAC4 [7] or to detect faulty beam position monitors in the LHC [8]. Deep neural networks were used to obtain predictions [9] and its uncertainties [10] in diagnostics for measuring beam properties at SLAC National Lab. At the University of Florida in Gainesville, relevant physical parameters for calculating the critical temperature of new superconducting magnets were discovered [11] with machine learning. Furthermore, eight different superconducting rf faults were classified with high accuracy at Jefferson Laboratory [12] using classic machine learning. However, to the best of our knowledge, none of the stated methods analyzed the parameters of the trained machine learning models, i.e., used explainable-AI, to explore the physical properties of the underlying phenomena. This is particularly relevant when making predictions that have a potential impact on machine protection and machine availability.

Overall, the objective of this work is to (1) analyze historical data of CLIC rf cavities with explainable-AI to better understand the behavior of breakdowns and to (2) investigate possibilities of data-driven algorithms for conditioning and operation of rf cavities.

The paper is organized as follows: Following this Introduction, Sec. II describes the experimental setup and data sources. Section III describes the methodology for data-driven modeling and gives insights into the design choices made, based on the characteristics of the available historical data. We further provide a comprehensive overview of rf-cavity breakdowns, convolutional neural networks for time series, and explainable-AI techniques. We then present the modeling and experimental results for two different data types, i.e., trend data in Sec. IV and event data in Sec. V. With explainable AI, we state that a pressure rise is the first sign of a breakdown and validate it empirically. The strengths and the limitations of our methodology are discussed, together with an outlook for possible future work in Sec. VI. Finally, we conclude our research in Sec. VII.

The code of our machine learning framework is publicly available.[1]

[1]https://github.com/cobermai/rfstudies.

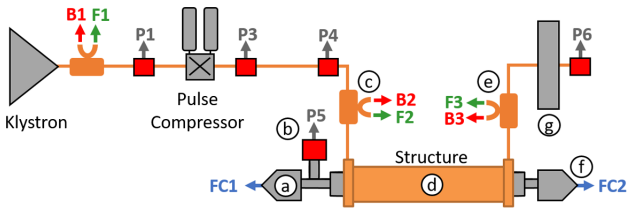


FIG. 1. Schematic of CERN's XBOX2 test stand. The red and green arrows show where the backward reflected traveling wave (B) and the forward traveling wave (F) rf signals are measured via directional couplers. The upstream and downstream Faraday cup signals are labeled FC1 and FC2. The locations of the ion pumps throughout the system are also shown (P). The lowercase letters mark the items also shown in Fig. 2.

II. EXPERIMENTAL SETUP

To investigate the challenges associated with the high-gradient operation and to validate the novel 12-GHz rf components for the CLIC project, CERN has commissioned three *X*-band klystron-based test stands named XBOX1, XBOX2, and XBOX3, respectively [13]. The test stands have been previously reported in detail [4,13]. To allow for better readability of this paper, we provide a short introduction to their structure and operation modes. While all three test stands are built with the same arrangement, they mainly vary depending on the specific components used. A schematic of the high-power portion of the XBOX2 test stand is shown in Fig. 1. The locations, denoted with lowercase letters, are also shown in a photograph of one of the test stands in Fig. 2. In each test stand, a 12-GHz phase-modulated low-level radio frequency (LLRF) signal is amplified to the kilowatt level and used to drive a klystron. The high-power rf signal produced by the klystron is then directed through a waveguide network to the rf cavity. To increase the peak power capability, each test stand is also equipped with

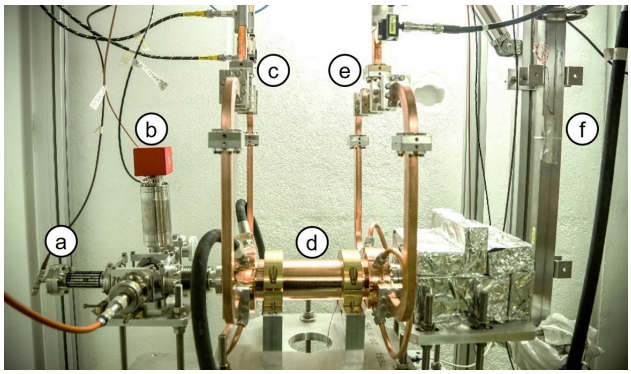


FIG. 2. Picture of a prototype accelerating structure installed in one of the test stands [16]. Visible are the upstream Faraday cup (a), an ion pump (b), the rf input (c) and output (e), the rf cavity under test (d), the shielded lead enclosure (f), and the high-power rf load (g).

specially designed energy storage cavities, also known as pulse compressors [14,15].

During operation, the forward (F) and backward (B) traveling rf signals are monitored via directional couplers. The gradient throughout the waveguide network is measured by directional couplers and logged by the control system. The XBOX2 and XBOX3 test stands are situated in a facility without beam capability. However, during high-field operation, electrons are emitted from the cavity surface and accelerated. This phenomenon, which is undesired in real operation, is known as dark current [17–19]. Monitoring the emitted current during operation is an important measure used in detecting cavity breakdowns, as will be shown later. During the operation of the test stand, the dark current is measured via two Faraday cups, situated on the structure extremities in the upstream (FC1) and the downstream (FC2) directions. Finally, the internal pressure is maintained and measured with a series of ion pumps (P) located throughout the waveguide network.

In Fig. 2, a prototype of the CLIC accelerating structure (d) is visible with the waveguide input (c) and output (e). The directional couplers and coaxial cables, which measure the high-power rf signals, can be seen at the top center, above these waveguide parts. The upstream Faraday cup (a), an ion pump (b), and the high-power rf load (g) are also visible. The downstream Faraday cup is situated inside a shielded lead enclosure (f) which is necessary for protection against the dark current.

Figure 3 shows two examples of different events, measured by the directional couplers and the Faraday cups. On the left side, the data from a healthy event are shown, and on the right side, a breakdown event is plotted. Figure 3(a) shows the approximately rectangular klystron pulse (F1). As is visible in Fig. 1, the test slot is equipped with a pulse compressor. To operate this device, phase modulation is applied to the klystron pulse, beginning after approximately 1700 samples of F1. Note that the position of the edge is not always at the exact position, as it can be changed by the operator without changing the performance of the system. Figure 3(b) shows the resulting "compressed" pulse which is delivered to the structure (F2). The device consists of two narrowband energy storage cavities linked via a hybrid coupler. As a consequence, upon receipt of the klystron pulse, most of the power is initially reflected, resulting in the sharp edge visible after approximately 200 samples (0.125 μs) of F2. As the storage cavities slowly begin to fill with energy and emit a wave, interference between the reflected and emitted waves occurs, resulting in the gradual change of amplitude in the transmitted waveform. When the phase of the incoming klystron pulse is modulated after approximately 1700 samples (1.0625 μs) of F2, the reflected and emitted waves constructively interfere, producing a short, high-power region that is flat in amplitude. Following the cessation of the klystron pulse, the remaining energy in the cavities is emitted, resulting in a gradual decay in the amplitude of the transmitted waveform. Further details on the design and operation of the pulse compressor are available in [20].

The signal which is reflected from the structure (B2) is shown in Fig. 3(c). As the accelerating structures are of the traveling wave design, nominally, the reflected signal is small. During breakdown events, however, the arc effectively acts as a short circuit, reflecting the incoming wave as shown on the right of Fig. 3(c). Fig. 3(d) shows the transmitted signal (F3). During normal pulses, this waveform is similar to the signal at the structure's input, while truncation is observed during breakdown events as most of the power is reflected back toward the input [see on the right of Fig. 3(d)]. Finally, the upstream and downstream Faraday cup signals are shown in Figs. 3(e) and 3(f), respectively.

All XBOX2 data are shown in Fig. 4. Specifically, the maximal value and the pulse width of the F2 signal with respect to the cumulative pulses for all data in 2018 are shown. Additionally, the cumulative breakdown count is shown. Initially, many breakdowns occur during the first part of the conditioning. Here, both the F2 maximal value and the pulse width value vary. The yellow area represents pulses, during which these F2 values were stable. These pulses will be used for further processing in Sec. III A.

A. rf cavity breakdowns

In high-gradient rf cavities, small surface deformations can cause a local enhancement of the surface electric field, resulting in substantial field emission and occasional plasma formation, i.e., arcing, which can damage the surface as shown in Fig. 5. The plasma which forms in the cavity during such breakdown events constitutes a significant impedance mismatch that reflects the incoming rf power.

Additionally, breakdowns are accompanied by a burst of current, which is generally a reliable indicator for structure breakdowns [18,22,23]. Minor fluctuations, which do not lead to the formation of plasma and the subsequent reflection of the incoming power detected by the Faraday cups, are defined as *activity* on the surface of the structure. In the XBOX test stands, these are measured by Faraday cups to reliably detect breakdowns and regulate the conditioning process (see Fig. 2 FC1 and FC2) [3,24]. Typically, at an accelerating gradient of 100 MV/m, Faraday cup signals of the order of 1 mA are observed in the test stands [18]. The threshold for structure breakdowns is typically set to 81.3% of the maximal resolution of the analog to digital converter in the Faraday cups, e.g., −0.615 to 0.615 V for XBOX2, which corresponds to currents in the hundreds of milliamps range. In Fig. 3, it is shown that during breakdown events, a large dark current is emitted, and thus the threshold on the Faraday cup signal

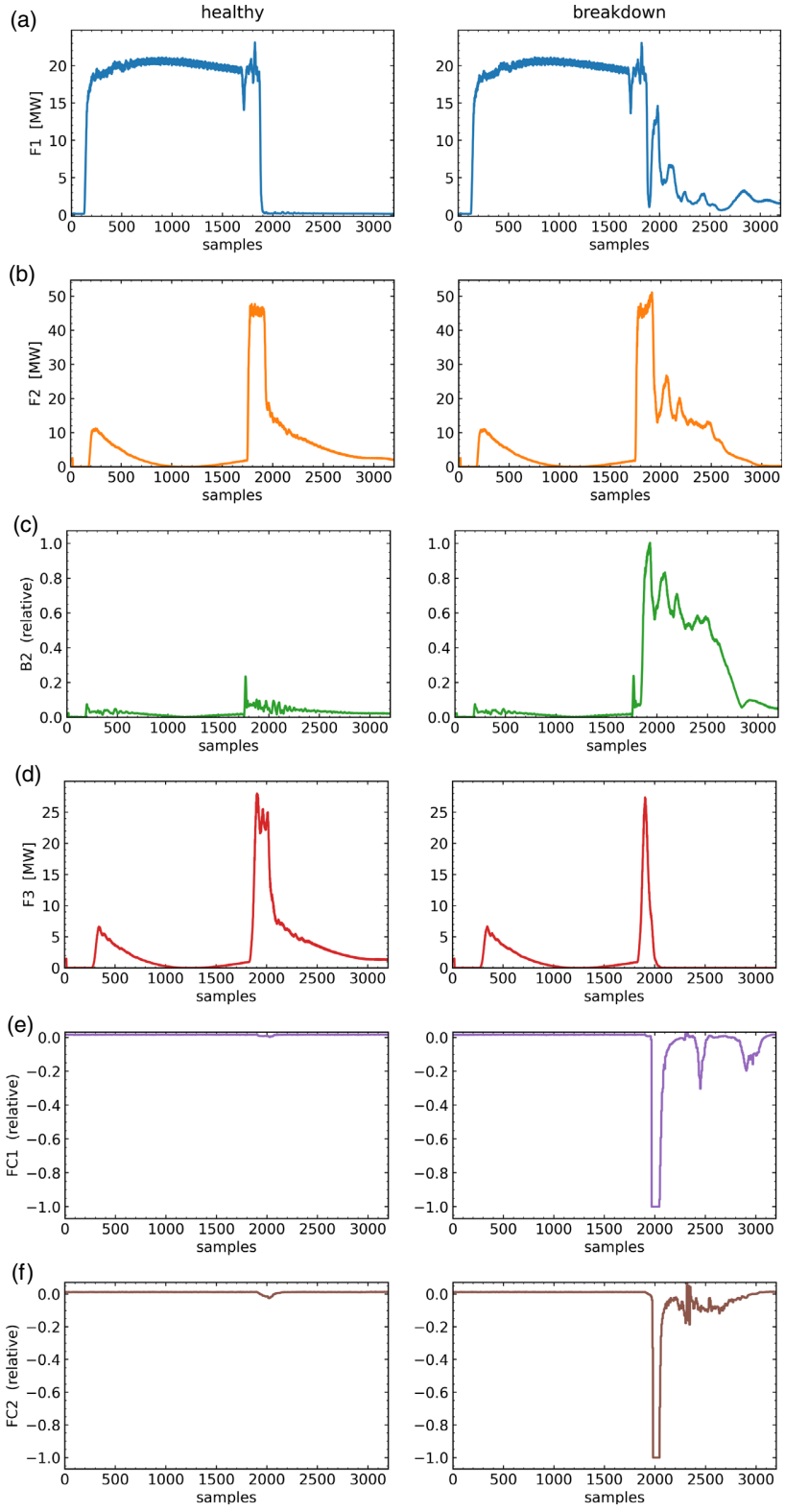


FIG. 3. Two examples of different events, showing the F1, F2, B2, F3, FC1, and FC2 signals. The left plots represent the signals of a healthy event, the right plots represent the signals of a breakdown event. All signals are 2 μs long. Note that the power amplitude of the forward traveling waves after the klystron (a), before the structure (b), and after the structure (d), are shown in MW. The power amplitude in the backward traveling wave (c), the upstream (e), and downstream (f) Faraday cup signals are shown relative to their maximum value, as no calibration coefficients were provided by the system.

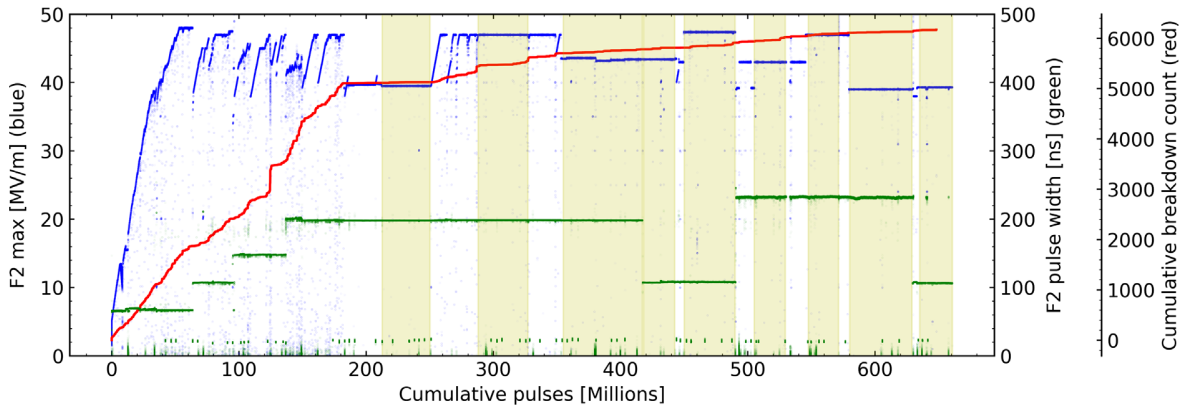


FIG. 4. Overview of the conditioning period, containing all data analyzed. The yellow area represents the runs during which the operational settings were kept stable and which we used for analysis. Additionally, the maximum power amplitude of the forward traveling wave signal F2 (blue), its pulse width (green), and the cumulative breakdown count (red) is shown.

(FC1, FC2) is well suited to distinguishing between healthy and breakdown signals.

Breakdowns usually occur in groups. When a breakdown is detected in the XBOX test stand, the operation is stopped for a few seconds. Afterward, operation is resumed by ramping up the input power within less than a minute.

During conditioning, the total number of breakdowns varies widely on the tested structure, which is why structures are generally more comparable in terms of the cumulative number of rf pulses. As a result, it has previously been proposed that conditioning proceeds primarily on the number of pulses and not solely on breakdowns [25]. This also aligns with the results of high-voltage dc electrode tests, where conditioning has been linked to a process of microstructural hardening caused by the stress associated with the applied electric field [26]. In addition to the copper hardness, the total number of accrued breakdowns is thought to be affected by the copper purity, the cleanliness of the structure [27] defined by the amount of dust and other contamination, the design of the cavity, and the level to which the cavity must be conditioned dependent on the nominal operating power and pulse length.

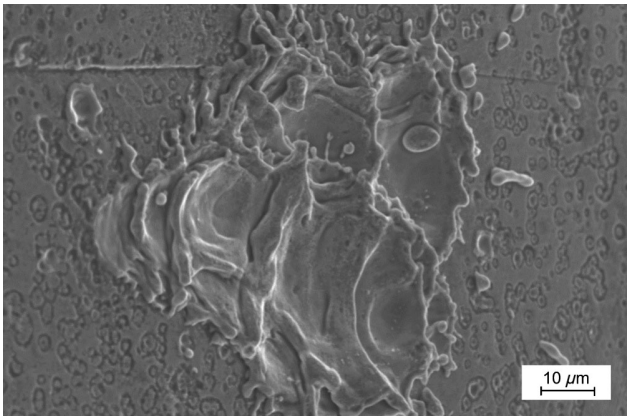


FIG. 5. Example of a crater after a breakdown on the surface of a copper rf cavity [21].

B. Data from experimental setup

90 GB of data from a period of 6 months in 2018 were produced during the operation of the XBOX2 test stand. The high-gradient cavity, tested during this time, was produced at the Paul Scherrer Institute in Switzerland [16,28]. The data are divided into so-called *trend* data and *event* data. Trend data contain 30 single scalar values, e.g., pressure measurements, temperature measurements, and other system relevant features. Event data contain six time-series signals of 2 μs length, with up to 3200 samples (see Fig. 3).

Figure 6 shows an example of the trend and event data logging mechanism. In the test stand, event data are acquired every pulse at 50 Hz and trend data are acquired

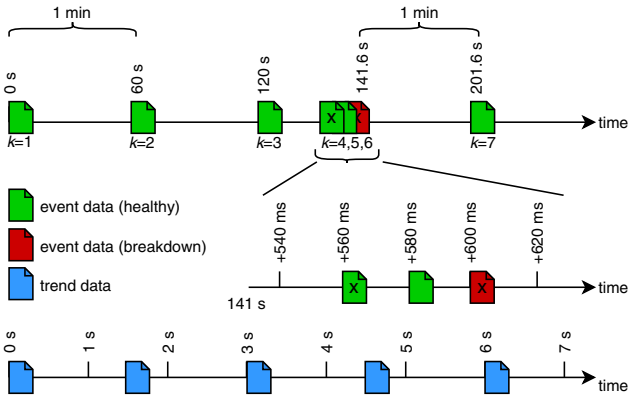


FIG. 6. Trend and event data logging. Event data signals are stored every minute during normal operation. In addition, in case a breakdown occurs, the breakdown event and the two prior healthy events are stored. Trend data features are stored every 1.5 s. The events indicated with an x are not considered for the prediction of breakdowns.

TABLE I. Information about different runs during which the operational setting was stable. Due to the limited amount of breakdowns, groups with similar F2 pulse width are formed for validation and testing during the modeling phase.

Run	No. of primary breakdowns	No. of follow-up breakdowns	F2 max (MV/m)	F2 pulse width (ns)	Group
1	10	3	35.8	182.4	Group 1
2	50	58	39.5	171.2	Group 2
3	41	38	34.6	161.5	Test
4	14	15	42.5	106.5	Group 3
5	35	62	42.7	100.8	Group 4
6	30	53	38.3	211.2	Group 5
7	21	16	37	186.1	Group 1
8	13	8	37.1	222	Group 5
9	5	7	34.9	102	Group 3

at up to 600 Hz. Due to the limited data storage of the experimental setup, the data cannot be stored with full resolution. The waveforms associated with an rf pulse are stored in an event data file every minute. In the case of breakdown events, the two prior rf pulses are logged in addition to the pulse, where the breakdown appeared. The corresponding trend data file is updated at a fixed rate every 1.5 s.

To go into more detail on the exact use of machine learning, we describe our data mathematically. Our data are a list of K-, M-dimensional multivariate time-series $\mathbf{X}_k = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ for $k \in \{1, \ldots, K\}$. Each of the M time-series has N samples, i.e., $\mathbf{x}_m \in \mathbb{R}^N$ for $m \in \{1, \ldots, M\}$. For both the event and the trend data, an event K is defined as an entry in the event data. The number of time-series M is given by the available signals of the power amplitude of the traveling waves and the Faraday cups for the event data. In the trend data, M is given by the number of available features, e.g., pressure, temperature, and other system relevant features. The number of samples N is defined by the number of samples in the event data signals and the amount of most recent data entries, of an event k in the trend data features.

Based on the Faraday cup threshold stated before, we assign a label *healthy* ($y_k = 1$) and *breakdown* ($y_k = 0$) to each event k. This results in a XBOX2 data set of shape $\{\boldsymbol{X}_k, y_k\}_{k=1}^K$. Using this notation, 124,505 healthy and 479 breakdown events were derived. We further define the first breakdown in each breakdown group as a *primary breakdown*, and all other breakdowns, within less than a minute of the previous breakdown, as *follow-up breakdowns*. With this definition, we split the given 479 breakdowns into 229 primary breakdowns and 250 follow-up breakdowns (see Table I). Compared to the high amount of healthy events, there is only a small amount of breakdown events. This so-called *class imbalance* is tackled by randomly sampling a subset of healthy events and by assigning class weights to the breakdown events during optimization of the algorithm and during the computation of the performance measure.

III. METHODOLOGY OF ANALYSIS

In this section, we discuss the background of the data processing used to generate the results. Generally, modeling schemes, for representing a system's behavior, are divided into model-driven approaches, where prior knowledge is embedded to represent a system's behavior, and data-driven approaches, where the system's behavior is derived from historical data. With the increasing amount of computational resources, available historical data, and successfully implemented machine learning algorithms, data-driven methods have become popular in many applications for failure prediction [29–31]. The choice of a data-driven algorithm is dependent on the application, the system complexity, and the amount of system knowledge available, as schematically shown in Fig. 7. The goal is to find the simplest model, which is capable to capture the relevant characteristics of the system under study [32].

When considering the goal of identifying a breakdown in an rf cavity, the most common approach relies on an expert setting a threshold [18] on a relevant quantity, e.g., the current measured by a Faraday cup, based on their knowledge about the system. An alternative approach could consider thresholds based on a statistical approach, which can be derived from the distribution of cavity breakdowns from past reliability studies [22]. However, such thresholds are not sufficient for highly nonlinear problems and complex system dependencies, like predicting rf breakdowns. In these cases, classical machine learning models, e.g., k-nearest neighbors (k-NN) [33], random forest [34], and support vector machine (SVM) [35], can be used to find these correlations and to derive optimal, more complex decision boundaries. In k-NN, an event is classified based on the majority class of its neighbors. Here, the neighbors are determined by finding the events with the closest Euclidean distance. A random forest is a combination of many decision trees to an ensemble. Decision trees learn simple decision rules, e.g., the FC1 signal reaches its saturation value, inferred from the most relevant characteristics of the problem, also called features. SVM on the other hand, learns a decision boundary that splits data into classes while maximizing the decision boundary margin. If features

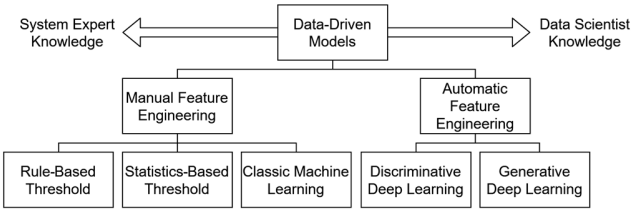


FIG. 7. Overview of different data-driven models.

in the data are not known *a priori*, deep learning [36], e.g., multilayer perceptrons, or convolutional neural networks, provides the ability to automatically extract and estimate them. Those methods are explained in detail in the modeling subsection. Deep learning can be categorized into discriminative deep learning, which directly models the output based on the input data, and generative deep learning, which models the distribution of the data from which the output is inferred. In order to develop an end-to-end time-series analysis framework without the necessity of manual feature calculations, we use deep learning models to analyze breakdowns in the CLIC rf cavities and show that they achieve superior results compared to classic machine learning approaches, such as k-NN, random forest, and SVM.

Specifically, we use discriminative deep learning models, due to their recent success to classify time-series signals [37]. By analyzing our models after training, we show how to extract system knowledge and physics insights, which then allows the extraction of models with reduced complexity.

For the labeled measurement data from the XBOX2 test stand, dedicated python toolboxes are used for feature calculation [38], time-series classification [37], and interpretation of model predictions [39]. Four steps of data processing and analysis, namely, transformation, exploration, modeling, and explanation, are carried out. These are detailed in the next paragraphs.

A. Transformation

Before training our machine learning models, we apply the following transformation steps to the data. All these steps contribute to fit the data and their properties to our models and include merging of event and trend data, filtering of unwanted events, and resampling and scaling of the event data signals.

Merging: Merging and synchronizing the trend data with the event data is a critical data transformation step to ensure the correct relative time order of the data (see Fig. 6). Particular caution is required to take the nearest past trend data samples for each event k.

Filtering: During our analysis, we only consider data during which the operational setting was stable, i.e., we filter the phases of commissioning or parameter adjustment. Specifically, we define so-called *runs* as the periods where the F2 max and F2 pulse width were kept constant. Table I shows the properties of the different runs, and Fig. 4 highlights these time periods in yellow. Due to the limited amount of breakdowns in certain runs and in order to increase the statistics, we also combine runs with a similar F2 pulse width (see Fig. 3) which we will use for modeling later on. Additionally, using a threshold of 650 kW on the power amplitude of the forward traveling wave signal F2, we further discard all events which only included noise, logged when the machine was actually not operating.

Scaling: The used features and signals have different units and different value ranges. To make them comparable, we standardize the data by subtracting the mean and dividing by the standard deviation. This way, all features and signals have a mean equal to 0 and a standard deviation equal to 1, independently of their units.

Resampling: In the event data, the Faraday cup signals (FC1, FC2) only have 500 samples compared to the 3200 samples from the other signals, as they are sampled with a lower frequency. Therefore, we interpolate the Faraday cup signals linearly to 1600 samples and selected only every second sample of the other signals.

B. Exploration

The goal of the exploration phase is to get an initial understanding of the event and trend data and to validate the transformation step. We compute 2D representations of the high dimensional data, in which each data point represents data of an event k, e.g., compressing all information that can be found in Fig. 6 on a 2D plane. This enables us to see correlations and clusters within the derived representations in a single visualization of the data. Outlier events, which are fundamentally different from other events, are further analyzed and, if applicable, neglected after further consultation with experts. Representation learning is a key field in machine learning with many methods available including but not limited to unsupervised machine learning methods like principal component analysis [40], stochastic neighbor embeddings [41], and representation learning methods based on neural networks [41–43].

In Fig. 8, we use two dimensional t-distributed stochastic neighbor embedding (2D-tSNE) [41], which converts pairs of data events to joint probabilities, i.e., the likelihood that they are similar. Close events have a high joint probability, and events far away have a low joint probability. Accordingly, 2D-tSNE creates representations in a 2D space and iteratively updates its location, such that the distributions P of the high-dimensional and the 2D space Q are similar. This equals the minimization of the Kullback-Leibler divergence [44] which measures the similarity between two distributions, i.e., $D_{\mathrm{KL}} = \sum_{x\in\mathcal{X}}(P||Q) = P(x)\log(\frac{P(x)}{Q(x)})$, where $\mathcal{X}$ is the domain of x.

After the dimension reduction, the different coloring of the representations is used to validate the steps of the transformation phase. No information about the coloring is given to the algorithm during training, which means that neither the runs nor the labels are used as input to compute the 2D-tSNE representations.

Figure 8 shows the 2D-tSNE dimension-reduced representation of the trend data during runs in which the operational settings were kept constant. The axis of the figure represents the two dimensions of the lower dimensional space, where correlations between the data samples

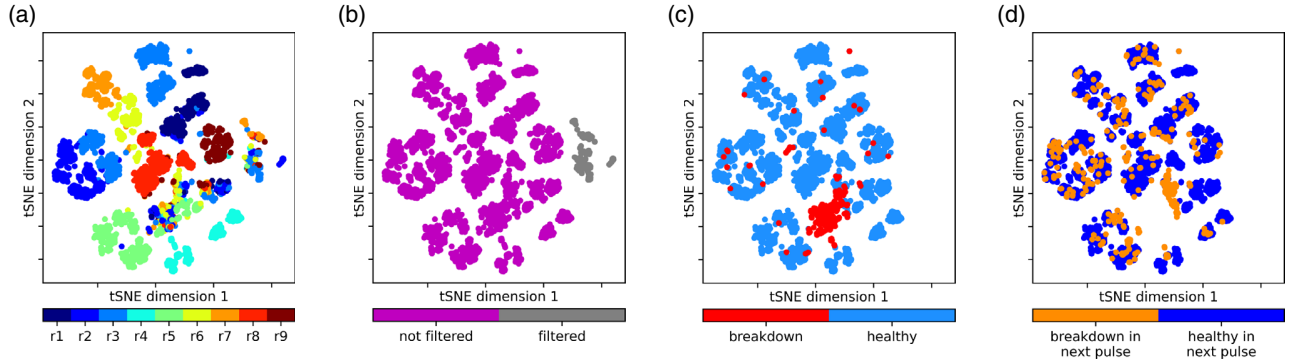


FIG. 8. 2D-tSNE of XBOX2 trend data during stable operation. The algorithm is able to distinguish between (a) stable runs, (b) not filtered and filtered events, and (c) breakdown and healthy events. In (d), no clear separation of events with a breakdown in the next pulse and a healthy event in the next pulse is possible. All representations in (c) are a subset of not filtered events in (b) and all representations in (d) are a subset of all healthy signals in (c).

are visible. First, representations are automatically colored, identifying the stable runs (a). This leads to clear clusters and validates the separation into different runs. In addition, two clusters with a mix of every run are formed. Their meaning becomes clear with different color schemes. The first cluster with mixed runs gets clear when using a coloring scheme as a result of the filtering in the transformation step (b), i.e., the filtering with the threshold on the power amplitude of the forward traveling wave signal F2.

Using all nonfiltered events from (b), we analyze if it is possible to classify breakdowns without giving the model any information about the label, i.e., if supervised modeling is necessary or if unsupervised learning would already be sufficient. Inspecting the clustering between breakdown and healthy events (c), it seems possible to use unsupervised learning for the classification, as many breakdown events form one cluster and are clearly separable from healthy events. This also explains one of the clusters of signals with mixed runs in (a).

As the unsupervised classification of breakdowns was successful, further investigations aim at identifying breakdowns during the following pulse, i.e., predicting breakdowns. Using all healthy events from (c), no clear unsupervised separation is possible for distinguishing events that are healthy in the next pulse from events that lead to a breakdown in the next pulse (d). Notably, the same phenomena can be observed when using other unsupervised methods, like autoencoders [42] or a higher dimensional space for clustering. As labels are available from the FC signals, we employ supervised learning techniques to distinguish the events shown in Fig. 8(d).

C. Modeling

The objective of the modeling phase is to find a function $f(\boldsymbol{X}_k)$ that predicts the output $\hat{y}_{k+1}$. This means that we classify whether a *breakdown in the next pulse* $\hat{y}_{k+1}$ will occur. This would be sufficient to protect the cavity and employ reactive measures to prevent its occurrence. The function $f(\boldsymbol{X}_k)$ is modeled with a neural network, and its parameters are optimized during training with the available historical data.

The results are obtained by discarding the event of the breakdown and the event two pulses before a breakdown, expressed with an x in the events $k = 4$, 6 in Fig. 6. This can be attributed to the fact that the equidistance of the event data is violated around a breakdown, which is corrected by this action. The network then solely focuses on using $\boldsymbol{X}_{k=5}$ to predict $y_{k=6}$.

1. Introduction to neural networks

To better understand the behavior of a neural network, we next give a brief overview of its structure. At a single neuron, a weight $w_{m,n}$ is assigned to each input $x_{m,n}$ of $\boldsymbol{X}_k := (x_{0,0}, \ldots, x_{M,N})$. The sum of the input multiplied by the weights is called the activation a of a neuron, which is further used as an input to an activation function $h(\cdot)$. This leads to the following equation:

$$f(\boldsymbol{X}_k) = h\left(\sum_m^M \sum_n^N w_{m,n} x_{m,n} + w_0\right), \tag{1}$$

where w_0 is a bias weight. Common activation functions are the sigmoid activation function $h(a) = 1/(1 + e^{-a})$ or the Rectified Linear Unit (RELU) $h(a) = \max(0, a)$. The choice of activation function depends on several factors [36], e.g., the speed of convergence and the difficulty to compute the derivative during weight optimization.

A neural network consists of several layers, where each layer includes several neurons which take the output of the previous layer neurons as an input. This allows the modeling of nonlinear properties in the data set. With a fully connected neural network, a neuron takes all outputs of the previous layer as an input, while in a convolutional neural network (CNN), the neuron only takes neighboring

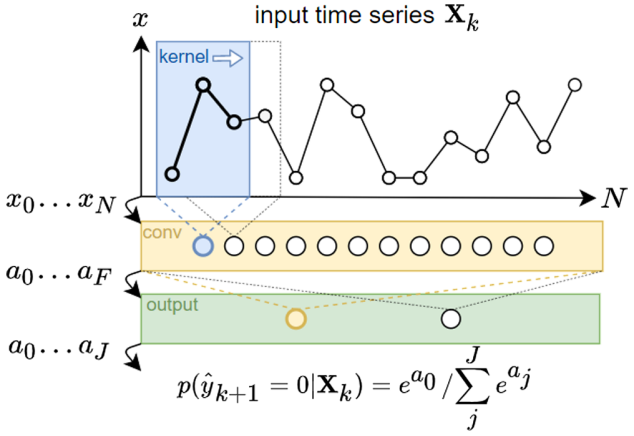


FIG. 9. Example of a convolutional neural network (CNN) for time-series prediction. For simplicity, the input $\boldsymbol{X}_k$ consists of only one signal, i.e., $m = 1$, and the network consists of only one hidden convolutional (conv) layer. As in most of our models, the softmax activation function is used as an output to derive $f(\boldsymbol{X}_k) = p(\hat{y}_{k+1}|\boldsymbol{X}_k)$ out of the activations a_j. In this example, the kernel size of the convolution layer is 3, the filter size is $F = 12$, and the probability of a breakdown in the next pulse ($y_{k+1} = 0$), is stated. In this case, the network would have 60 trainable weights.

neurons' output of the previous layer as an input. A CNN, therefore, creates correlations with neighboring inputs.

Essential parameters of a CNN are shown in a simple example in Fig. 9. The *kernel size*, defines the number of neighboring neurons used from the previous layer, and the *filter size*, defines the number of neurons in the current layer. The name filter is derived from the fact that a convolution can also be seen as a sliding filter over the input data. Furthermore, *pooling* refers to the method used for down-sampling a convolution to enhance the created correlations. Pooling can be either *local*, over each dimension separately, or *global*, over all dimensions. Two common pooling methods are *maximum pooling*, where the maximum of a window is taken as an output, and *average pooling*, where the mean of a window is taken as an output.

2. Learning of neural networks

Weight optimization is typically achieved with *gradient descent* methods using a loss function. For classification tasks with two classes, typically the cross-entropy-loss $E = -[y \log(p) + (1 - y) \log(1 - p)]$ is chosen, where y is the known class and p is the predicted class probability. In a process with i iterations, called *epochs*, a neuron's weight $w_{m,n}$ is then optimized by $w_{m,n}^{i+1} = w_{m,n}^{i} - \eta \nabla_w E$. Here, $\eta > 0$ is the learning rate, and $\nabla_w E$ is the gradient of the loss dependent on the weights. The gradient descent optimization can be further accelerated with more sophisticated optimizers. Specifically, we use the ADAM optimizer [45] in our models. It varies the learning rate dependent on the mean and the variance of the gradient. In Fig. 14(b), we visualize the learning process of our models, by showing the models' loss with respect to the epochs.

3. Advanced architectures

Due to their ability to learn correlations of neighboring inputs, CNNs contributed to the recent success of machine learning, finding many applications in image classifications [46], language processing [47], and time-series classification [37].

(i) *time-CNN*: The time CNN was originally proposed by Zhao *et al.* [48] and consists of two average pooling convolutional layers with 6 and 12 filters with a kernel of size 7. It uses the mean-squared error instead of the categorical cross-entropy-loss [44] for weight optimization, which is typically used in classification problems. Consequently, the output layer is a fully connected layer with a sigmoid activation function. Due to this architecture, the time-CNN has 4976 trainable weights and is therefore the model with the fewest parameters in our studies.

(ii) *FCN*: The fully convolutional network was originally proposed by Zhao *et al.* [49] and consists of three convolutional layers with 128, 256, and 128 filters of kernel size 8, 5, and 3. In each layer, batch normalization is applied, normalizing the output of the previous layer in each iteration of the weight optimization [50]. This leads to faster and more stable training. Each convolutional layer uses a RELU activation function, except the last one, where the output $a_1, \ldots, a_J$ is globally averaged and fed into a softmax activation function $h_i(a_1, \ldots, a_J) = e^{a_i} / \sum_j^J e^{a_j}$ to obtain the output probability $p(\hat{y}_{k+1}|\boldsymbol{X}_k)$ for $i = 1, \ldots, J$, where J is the number of different labels. The model has 271,106 trainable weights.

(iii) *FCN-dropout*: It is of similar architecture as the FCN with the same number of 271,106 trainable weights. In addition, it has two dropout layers after the second convolution and the global average pooling layers as proposed by Felsberger *et al.* [29]. This dropout layer is skipping neurons during training randomly with a probability of $p_{\text{drop}} = 0.5$, which improves the generalization of the model.

(iv) *Inception*: Inspired by the Inception-v4 network [51], an inception network for time-series classification has been developed [52]. The network consists of six different inception modules stacked to each other, leading to 434,498 trainable weights. Each inception model consists of a so-called *bottleneck layer*, which uses a sliding filter to reduce dimensionality and therefore avoids overfitting. Additionally, several filters are slided simultaneously over the same input and a maximum-pooling operation is combined with a bottleneck layer to make the model less prone to small perturbations.

(v) *ResNet*: The residual network was originally proposed by Zhao *et al.* [49] and consists of three residual blocks, i.e., a group of three convolutional layers. This architecture leads to 509,698 trainable weights. This relatively deep architecture

is enabled by using skip connections after each block. This skip connection is a shortcut over the whole block and provides an alternative path during weight optimization which reduces the risk of vanishing gradients [36]. The kernel size of the convolutional layers is set to 8, 5, and 3 in each residual block for the fixed number of 64 filters in each layer. The activation function, the batch normalization, and the output layers are similar to the FCN.

All models were trained on a single Nvidia Tesla V100 GPU. This took on average 24 min for the event data and 9 min for the trend data. Once the models were trained, one prediction took 27 ms for the event data and 18 ms for the trend data using TensorFlow [53] to compile the model without any optimization or compression. However, due to the random weight initialization and depending on the network, the training time slightly varied.

When using a softmax activation function in the last layer, the output of the function in Eq. (1) is the probability of the next event being healthy or a breakdown, i.e., $p(\hat{y}_{k+1}|\boldsymbol{X}_k) \in [0,1]$. To receive a binary label, $\hat{y}_{k+1} \in \{0,1\}$, it is necessary to set a threshold to the probability. The receiver operating characteristic (ROC) curve is a plot that shows how this threshold impacts the relative number of correctly classified labels as a function of the relative number of falsely classified labels. The ROC curve of the best models for each prediction task is shown in Fig. 14(a). We use the area under the ROC curve (AR) to rate the performance of our models. This is a measure of the classifier's performance and is often used in data sets with high class imbalance [54]. Intuitively, this score states the probability that a classifier designed for predicting healthy signals ranks a randomly chosen healthy event k^+ higher than a randomly chosen breakdown event k^-, i.e., $p[f(\boldsymbol{X}_{k^+}) > f(\boldsymbol{X}_{k^-})]$. An AR score of 1 corresponds to the classifier's ability to correctly separate all labels, while an AR score of 0 represents the wrong classification of all labels.

For training, validation, and testing of our model, we merged runs with similar F2 pulse width into groups as shown in Table I, as some runs have a small number of breakdowns. Specifically, we use *leave-one-out-cross-validation* on the groups. This means we iterate over all possible combinations of groups, while always leaving one group out for validation. After adjusting the model weights, e.g., the class weight, we then test our model on data from run 3.

The mean score AR_μ over all iterations and its standard deviation, AR_σ, are stated in the results together with the test result AR_t. In order to ensure that our model provides a good generalization to new data, we aim that AR_t of the test set should be within $\mathrm{AR}_\mu \pm 2\mathrm{AR}_\sigma$. To compare deep learning models with classic machine learning models, we additionally present the AR score of k-NN, random forest, and SVM algorithms. The hyperparameters of these models have been optimized during a sensitivity analysis. Specifically, we used $k = 5$ neighbors for k-NN, $t = 500$ decision trees in random forest, and the *radial basis function* for the SVM, with $C = 1, \gamma = 3.3 \times 10^{-2}$ for trend data and $C = 1, \gamma = 7.2 \times 10^{-5}$ for event data. For a detailed description of these hyperparameters, we refer to existing literature [44].

D. Explainable AI

To interpret the "black box" nature of deep neural networks, explainable AI recently gained attention in domains where a detailed understanding of the driving factors behind the results is of primary importance. In fields like medical applications [55,56], criminal justice [57], text analytics [58], particle accelerators [29], and other fields in the industry [59], experts cannot simply accept automatically generated predictions and are often even legally obliged to state the reasons for their decision. To reliably predict breakdowns in rf cavities, the explanation of a model is of similar importance. Hence, we utilize explainable AI in our studies to provide the experts with any relevant information used by the model to aid in interpreting the behavior of data-driven models, build trust in the prediction, validate the results, and find possible errors within the earlier data processing steps. Additionally, understanding why a prediction is made may shed light on the underlying physics of vacuum arcs and thus aid in future design decisions pertaining to high-gradient facilities.

Explainable AI is divided into event-wise explanation, where each prediction of the model is analyzed separately, and population-wise explanation, where all predictions are investigated at once. Event-wise explanation enables experts to gain trust in a specific prediction. The choice of event-wise explanation algorithms is dependent on the input, i.e., image, text, audio, or sensory data, and the preferred explanation technique, i.e., by showing the sample-importance [60] or by explanation-by-example [61]. Important samples are often computed with additive feature attribution methods [60,62,63], which calculate a local linear model for a given event to estimate the contribution of a feature to one prediction. Alternative gradient-based methods aim to determine the features that triggered the key activations within a model's weights [64,65]. Explanation-by-example states reference examples on which the prediction is made, by using the activation of the last hidden layer in a neural network and searching for similar activations of events in the training set [61].

Population-wise explanation helps experts to gain trust in the model and to select relevant input features for the predictions. In its simplest form, this is achieved with a greedy search [66], or deep feature selection [67] which applies similar techniques to regularized linear models [34,68]. However, both of the stated methods are very computationally intensive for deep learning models. A more efficient method proposes to train an additional selector network to predict the optimal subset of features for the main operator network [69].

In our studies, event-wise explanations are converted into population-wise explanations by looking at the distribution of a subset of event-wise explanations [70]. Our event-wise explanations are calculated with an additive feature attribution method [60]. This means we define a model

$$g(\boldsymbol{X}_k) = \sum_{m}^{M} \sum_{n}^{N} \phi_{m,n} x_{m,n} + \phi_0, \tag{2}$$

which is approximating the output $f(\boldsymbol{X}_k)$ for one event k, where $\boldsymbol{X}_k$ is either the trend data or the event data. In this local linear model, $\phi_{m,n}$ equals the contribution of the feature $x_{m,n}$ to the output $f(\boldsymbol{X}_k)$ and is called the *feature importance*. To calculate $\phi_{m,n}$, we assign a reference value to each neuron. This reference value is based on the average output of the neuron. When a new input value $x_{m,n}$ is fed into the network, a contribution score is assigned to the neuron, based on the difference between the new output and the reference output. All contribution scores are then back-propagated from the output to the input of the model f, based on the rules from cooperative game theory [71]. The contribution scores $\phi_{m,n}$ at the input are called SHapley Additive exPlanation (SHAP) values [39] and are used to explain our produced results.

This interpretation is, however, different for trend and event data. In trend data, the SHAP values are interpreted as feature importance, stating the contribution of, e.g., the pressure to the prediction of breakdowns. In event data, the SHAP values are given for each time-series sample, e.g., the importance of each of the 3200 samples in the F1 signal. Here, the mean of all SHAP values in one signal is taken to derive the overall importance of a signal.

IV. RESULTS USING TREND DATA

In this section, we report the results of applying the methodology of analysis described above, using the trend data of the XBOX2 test stand. Specifically, we use the $N = 3$ closest trend data point in the past, of an event k, as described in Sec. II B. Each trend data event consists of $M = 30$ values, including pressure, temperature, and other system relevant features, measured in the test stand.

A. Modeling

Table II shows the AR score for the prediction of breakdowns with trend data. The results of the different model types described in the previous section are reported for comparison and discussed in detail. For each type of breakdown, the best model score is highlighted in bold. We chose the best model based on four decision criteria: (i) the average performance of the model AR_μ, (ii) the ability of the model to generalize within runs $\mathrm{AR}_\mu \pm 2\mathrm{AR}_\sigma$, and (iii) the ability of the model to generalize to new data AR_t. Additionally, we consider (4) the simplicity of the model given by the number of trainable weights and the complexity of the model structure, as this has a direct impact on the computational cost, which we want to minimize.

The ResNet model is able to predict primary breakdowns with an average AR score of 87.9%. With 7.2%, the standard deviation is much higher compared to the prediction of follow-up breakdowns, but still, the best generalization capability compared to the other models for predicting primary breakdowns. The inception network scores best on the test set with 82.9%. However, since the ResNet model performs best on two out of four decision criteria, we consider it the best for predicting primary breakdowns.

The relatively high standard deviation in the prediction of primary breakdowns states that the patterns learned by the network vary, i.e., the indicators of a primary breakdown differ dependent on the runs on which the network is trained.

With an AR_μ score of 98.7% and an AR_t score of 98.6%, the inception model predicts follow-up breakdowns best. This means that for the training set, there is a probability of 98.7% that our model assigns a higher breakdown probability to a randomly chosen breakdown event than it assigns to a randomly chosen healthy event. The score is 0.1% less when the model uses the test data. This indicates

TABLE II. AR score of different models, predicting primary, follow-up, and all breakdowns with trend data. The model for each column is highlighted in bold. AR_μ relates to the average AR score of different validation sets and AR_σ to the standard deviation. The trained model is finally tested on the test set with a performance AR_t.

	(1) Primary breakdowns			(2) Follow-up breakdowns			(3) All breakdowns		
Model	AR_μ (%)	AR_σ (%)	AR_t (%)	AR_μ (%)	AR_σ (%)	AR_t (%)	AR_μ (%)	AR_σ (%)	AR_t (%)
k-NN	61.0	7.4	63.1	89.8	8.1	92.9	76.7	8.0	75.9
SVM	68.8	10.0	73.8	93.6	5.7	97.0	84.2	9.8	87.8
Random forest	81.0	16.7	82.5	96.9	3.5	96.5	87.9	13.3	90.0
Time-CNN	55.2	11.0	48.1	92.8	3.8	87.6	67.7	6.3	66.0
FCN	86.1	8.7	81.0	98.2	1.0	97.8	**93.8**	**4.2**	**90.6**
FCN-dropout	84.9	9.0	81.7	95.6	3.0	97.3	92.7	4.6	90.6
Inception	85.4	8.5	82.9	**98.7**	**1.6**	**98.6**	92.3	4.8	90.9
ResNet	**87.9**	**7.2**	**80.4**	98.7	1.4	98.0	93.1	4.6	90.1

that the model generalizes well to new data, as the AR_t score is within AR_σ. The ResNet model offers similar results and an even smaller AR_σ. However, the inception model is preferred for the prediction of follow-up breakdowns due to its fewer trainable weights.

Looking at the prediction of both follow-up and primary breakdowns, the AR scores are approximately averaged compared to the two separate AR scores, the number of primary and follow-up breakdowns is similar. This indicates that the model finds similar patterns for both breakdown types. Here the FCN model scores best with an AR_μ score of 93.8% and an AR_σ of 4.2%. While the AR_t score of 90.6% is slightly lower than in the inception model, the FCN model has significantly fewer trainable weights.

The time-CNN model generally performs poorly compared to the others. A possible reason for this is that the low amount of trainable time-CNN weights cannot capture the complexity of the data. Additionally, the structure of the model might be insufficient. Here, we specifically refer to the unusual choice of Zhao *et al.* [48] to select the mean-squared error and not the cross-entropy-loss. The mean-squared error is typically used in regression problems, where the distribution of data is assumed to be Gaussian. However, in binary classification problems, the data underlie a Bernoulli distribution, which generally leads to better performance and faster training of models trained with the cross-entropy-loss [72]. The lower performance of the time CNN suggests that the mean-squared error should not be used in classification tasks for XBOX2 breakdown prediction.

Random forest is the only classic machine learning algorithm that achieves similar AR_μ and AR_t scores compared to deep learning. For example, when looking at the prediction of primary breakdowns, the AR_t score of 82.5% is even higher than the ResNet score of 80.4%. However, the standard deviation AR_σ of 16.7% is more than twice as high compared to the ResNet model, which makes its prediction less reliable. The higher standard deviation of classic machine learning compared to deep learning is also observed in the other breakdown prediction tasks.

For each prediction task, the ROC curve of the best model's test set performance is shown in Fig. 14(a). Here, the true positive rate corresponds to the percentage of correctly predicted healthy events, and the false positive rate corresponds to the amount of falsely predicted healthy events. For predicting primary breakdowns, the ResNet ROC curve (1) is plotted in green. Note that the AR_t score, corresponding to the area under the ROC curve, is 80.4% in this case. One can see a slow rise, which reaches a true positive rate of 1.0 at a false positive rate of about 0.4. For predicting follow-up breakdowns, the inception model (2, red) has the highest $AR_t = 98.6\%$ which is confirmed by the large area under the red curve. The curve of the FCN (3, blue) for predicting all breakdowns with $AR_t = 90.6\%$, is a mixture of the primary and follow-up breakdown prediction curves. It is reaching a true positive rate of 1.0 at a false positive rate of about 0.2. Using this information, it can be decided at which probability $p(\hat{y}_{k+1} = 1|\boldsymbol{X}_k)$ an event should be classified as a healthy event. Considering the inception model (2, red) for predicting follow-up breakdowns, a good choice would be the "edge," where the true positive rate is ∼1 and the false positive rate is 0.05. Here, almost all healthy events are labeled correct, while 5% of all breakdowns are falsely considered to be healthy events. However, the final choice of the probability threshold depends on the final application setting of the model and the consequences of false positives and false negatives, further discussed in Sec. VI.

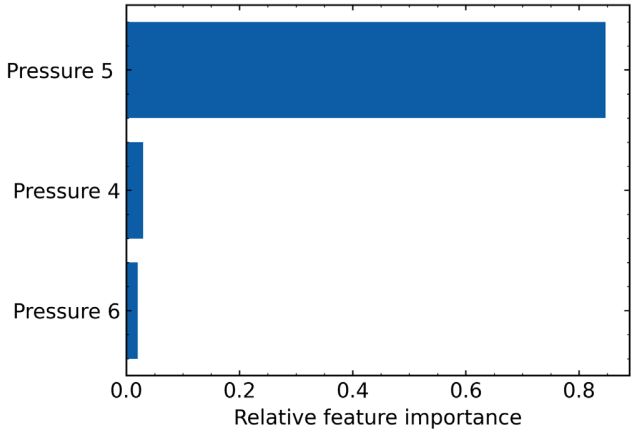


FIG. 10. The three most important trend data features, selected from 30 features in total, for predicting primary breakdowns with trend data.

B. Explainable AI

As primary breakdowns are generally considered a stochastic process [73], the good performance in Table II on predicting primary breakdowns is especially interesting. Hence, we focus on the trained models to gain deeper insights into the reason behind the good prediction results.

Figure 10 shows the importance of the features $\boldsymbol{X}_k$ for the prediction of primary breakdowns with trend data. Pressure 5 measurements, indicated also with P5 in Fig. 1, is the most relevant feature by a very significant margin, even when compared to the second and third most relevant features. By looking at this signal in more detail, for the different breakdown events in Fig. 11, it can be seen that the highest pressure reading is logged up to a few seconds before a breakdown event. Initially, it was expected that the pressure should be highest after the breakdown is detected via the Faraday cups, after the arc formation and the burst of current. However, here we observe the peak value beforehand.

We investigated the possibility that the observed effect is caused by a systematic error or a timing misalignment in

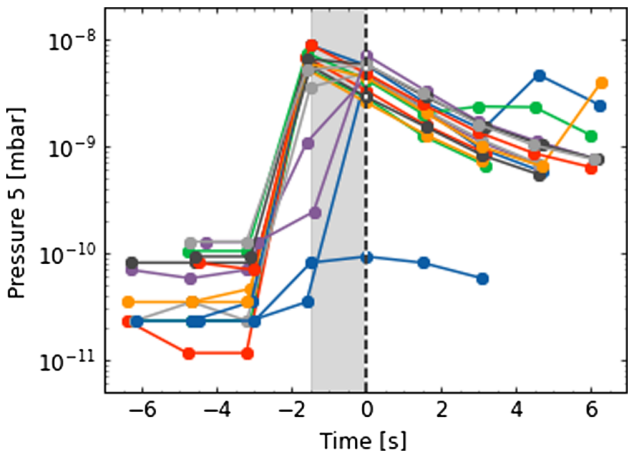


FIG. 11. Data samples of pressure 5, aligned to the interlock state of the test stand. The gray area represents the confidence interval, i.e., the window of time covering the previous 75 pulses in which the breakdown occurred. Data indicate that the pressure begins to rise before an interlock is triggered with the Faraday cup and the reflected traveling wave signals.

pressure rise, which could have occurred due to the logging algorithm in the control software of the XBOX2 test stand. We utilized a trend data feature of the XBOX2 test stand, which indicates whether the test stand was in an interlocked state, i.e., pulsing is inhibited, or if it is pulsing. Notably, this feature was not used for prediction. Since the pulse rate is 50 Hz, we know that the breakdown must have occurred in 1 of the 75 pulses prior to the interlock. Figure 11 shows the trend data features of the internal beam-upstream pressure during run 4. All data are aligned to the interlock time of the mentioned XBOX2 feature, which is indicated with the black dashed line. The gray area is the confidence interval, covering the previous 75 pulses during which a breakdown occurred, and the interlock signal was generated. A rise in pressure is visible in all data samples before the interlock is triggered. However, the low trend data sampling frequency means significant aliasing is possible, and so the true peak pressure could occur either earlier or later than is shown in the data. Therefore, the internal beam-upstream pressure signal should further be investigated.

Notably, during breakdowns, the vacuum readings located the furthest away from the structure demonstrated a markedly smaller rise which occurred later in time than that observed in the pumps located closest to the structure. This aligns with the expectation that the pumps situated farthest from the site of a given pressure change should measure it last due to the vacuum conductivity of the waveguide.

Generally, significant outgassing is observed in the early stages of component tests in the high-gradient test stands, and a conditioning algorithm that monitors the vacuum level and regulates the power to maintain an approximately constant pressure has been designed specifically for this early phase of testing [13]. It is known, that the exposure of fresh, unconditioned surfaces to high-electric fields results in measurable vacuum activity, however, it is unclear why a measurable pressure rise may occur prior to breakdown when a stable high-gradient operation has been reached. One potential explanation is that the phenomenon may be related to the plastic behavior of metal under high fields. In recent years, it has been proposed that the movement of glissile dislocations, which is a mobile dislocation within the metal, may nucleate breakdowns if they evolve into a surface protrusion [74]. If such dislocations migrate to the surface, then the previously unexposed copper may act as a source for outgassing, resulting in measurable vacuum activity while also being liable to nucleate a breakdown soon thereafter.

C. Experimental validation

To experimentally validate the phenomenon of the pressure rise before the appearance of a breakdown in the XBOX2 test stand, a dedicated experiment was conducted on a similar rf cavity in the XBOX3 test stand. In case of a substantial pressure increase which may indicate a vacuum leak, klystron operation is inhibited and thus no further high-power rf pulses can be sent to the structure. To facilitate interlocking, the pumps throughout the waveguide network are checked at 600 Hz, several hundred Hz higher than the rf repetition rate. However, due to the limited storage space, not all data are logged (see Fig. 6).

If the pressure begins to rise several pulses prior to a breakdown event, then by appropriately setting the threshold, it is possible to generate an interlock signal and stop pulsing prior to the breakdown. If the rise in pressure is caused by the start of processes that lead to a breakdown then by resetting the interlock and resuming high-field operation, it is assumed that the processes may continue, and a breakdown will then occur shortly after the initial interlock was generated. To validate this hypothesis, a 3-h test slot was granted in CERN's XBOX3 test stand during which the threshold for vacuum interlocks was set to be abnormally low, close to the pressure, at which the test stands generally operate. During this time slot, the data in Fig. 12 was recorded. The procedure of the experiment is visualized in Fig. 13. After detecting the early pressure rise with explainable AI, this finding allows us to simply use a threshold above 10% of the nominal pressure (see Fig. 11). Naturally, a large sample size, i.e., number of primary breakdowns, is desirable to validate the phenomenon. The breakdown rate may be considerably increased by raising the operating gradient although, as shown in Fig. 11, the pressure remains considerably elevated following breakdown events, necessitating a recovery period of several minutes before the pressure returns to the prebreakdown baseline. Additionally, increases in power are associated with increased vacuum activity and so stable, low pressure operation was favored throughout the run to avoid false

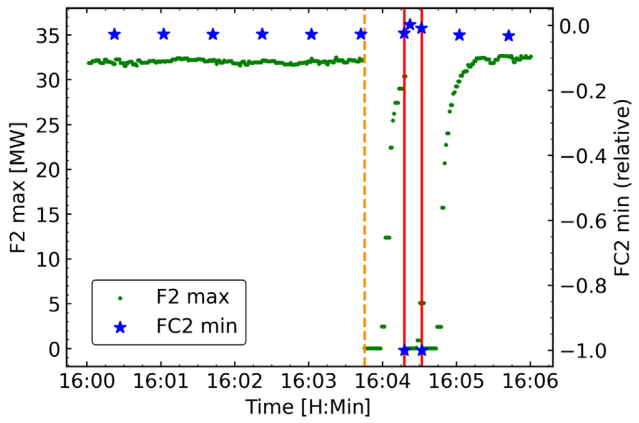


FIG. 12 Maximum value of the structure input power amplitude of the forward traveling wave (F2 max) and minimal value of the downstream Faraday cup signal (FC2 min) during the experiment to predict breakdowns. The orange dashed line shows an interlock, activated by a threshold on the pressure signal, meant to prevent a breakdown. The maximum structure input power amplitude of the forward traveling wave is logged as a feature in the trend data every 1.5 seconds. The minimal value of the downstream Faraday cup signal is extracted from the event data according to Fig. 6.

alarms and ensure reliable interlocking. During the 3-h experiment period, five primary breakdowns occurred, two of which were preceded by a vacuum interlock. One such example is shown in Fig. 12.

In Fig. 12, an interlock was produced and then reset several seconds later. The reset was done by removing the interlock thresholds temporarily to allow the test stand to ramp back up to nominal conditions and resume high-power operation. After ramping up in power, two primary breakdowns occurred, as shown by the red lines.

These instances align with what was observed in the historical data. However, given the relatively few primary breakdowns, further experiments are necessary. To overcome the alignment and resolution issues present in the historical data, an improved test stand logging system is currently being developed to record pressure as event data with high resolution.

V. RESULTS USING EVENT DATA

In this section, we report the results of applying the methodology of the analysis described above, using only the event data of the XBOX2, as shown in Fig. 3. We report these results separately to show that our models do not solely rely on the pressure reading as described in the previous section to successfully predict breakdowns.

A. Modeling

In Table III, we summarize the results of predicting breakdowns with event data based on the models described in Sec. III. We use the same decision criteria as in the previous Sec. IV A to select the best model.

With a mean validation score of 56.6% and a test score of 54.0%, the FCN-dropout performs best on the prediction of primary breakdowns. Although the AR_σ score of 8.3% is higher than in the inception model, the FCN-dropout model is preferred since it has significantly fewer trainable weights. Note that a score of 50% equals a random classifier, which guesses the output. Despite the stochastic behavior of primary breakdowns, our models exceed the expected 50%. However, the result is significantly lower compared to the prediction of primary breakdowns with trend data in Table II. This shows that the pressure rise found in analyzing the trend data is the main indicator for predicting primary breakdowns, given the available data and the described models.

Nevertheless, using event data, the models accurately predict follow-up breakdowns. Here the FCN model is preferred with an AR score of 89.7% for the prediction of follow-up breakdowns and shows the best generalization result on the test set with 91.1%. The AR score of 89.7% implies that with a probability of 89.7%, the FCN model attributes a higher breakdown probability to a randomly selected breakdown event than a randomly selected healthy event. The FCN-dropout offers better generalization on

TABLE III. AR score of different models, predicting primary, follow-up, and all breakdowns with event data. The model for each column is highlighted in bold. AR_μ relates to the average AR score of different validation sets and AR_σ to the standard deviation. The trained model is finally tested on the test set with a performance AR_t.

	(4) Primary breakdowns			(5) Follow-up breakdowns			(6) All breakdowns		
Model	AR_μ (%)	AR_σ (%)	AR_t (%)	AR_μ (%)	AR_σ (%)	AR_t (%)	AR_μ (%)	AR_σ (%)	AR_t (%)
k-NN	49.6	1.2	48.4	61.4	10.1	58.7	57.2	10.0	54.9
SVM	50.0	0.0	50.0	63.0	7.8	62.5	57.3	3.6	56.3
Random forest	48.2	3.4	50.0	66.9	9.2	73.0	58.4	6.9	59.7
time-CNN	52.7	3.4	51.9	79.2	12.8	82.1	59.8	7.7	66.6
FCN	54.7	9.8	52.8	**89.7**	**8.1**	**91.1**	66.8	12.5	68.7
FCN-dropout	**56.6**	**8.3**	**54.0**	89.1	5.3	83.7	**65.2**	**7.3**	**67.3**
Inception	52.6	3.6	49.9	87.9	8.4	90.5	65.9	13.6	67.1
ResNet	51.9	7.0	53.5	88.7	7.7	89.9	67.2	14.3	68.5

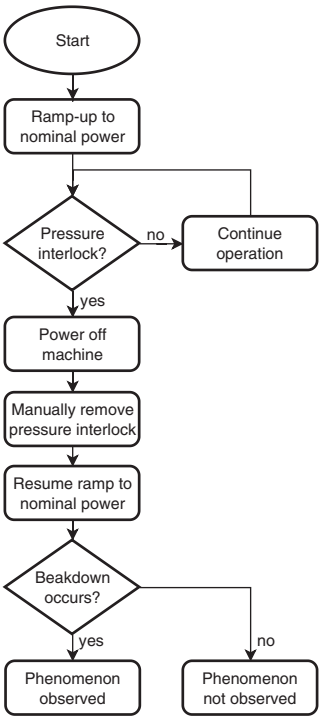


FIG. 13. Flowchart showing the procedure of the experiment. The pressure interlock was set to 10% above a nominal pressure. The Faraday cup signals and the reflected traveling waves were used to detect the breakdown.

different runs with an AR_σ of 5.3%, but relatively bad generalization on the test set with an AR_t score of 8.7%. The inception model and the ResNet model archive similar results, but utilize more trainable weights, which is disadvantageous.

With 8.1%, the standard deviation of predicting follow-up breakdowns with event data is much higher than the prediction of follow-up breakdowns with trend data in Table II. This means that the patterns learned by the network vary more when our models are trained on event data than on trend data. The values in Table I underline this conclusion, as the F2 max values and the F2 pulse width values are different depending on the run. The influence of the F2 max deviation is mitigated by the standardization of each signal by its own mean. However, the fluctuation of the F2 pulse width values makes it harder for the network to find common patterns in the time-series signals. In the trend data, the model mainly focused on the pressure rise, which is a phenomenon occurring across all runs.

Like in Table II, the mean of both primary and secondary breakdown prediction scores is close to the prediction of all breakdowns. This again indicates that the patterns detected are used for both follow-up and primary breakdowns. However, in primary breakdowns, this pattern occurs only rarely, leading to lower performance compared to the prediction of breakdowns with trend data. Here, the ResNet model has the best AR_μ score with 67.2%, the FCN-dropout model has the best AR_σ score of 7.3%, and the FCN model has the best AR_t score with 68.7%. Overall, the FCN-dropout model is considered best, due to the significantly lower standard deviation and the relatively low amount of trainable weights compared to the inception model.

In contrast to the trend data results in Table II, all classic machine learning methods show lower performance than the deep learning models. Figure 7 shows that classic machine learning requires features as input. When those features are given, as they are in the trend data, similar performance to deep learning is achieved. However, in the event data, time-series signals are used as input instead of features. Classic machine learning models are not able to generalize well anymore. Deep learning models automatically determine features in their first layers, and therefore, reach higher performance in all three prediction tasks.

Figure 14(a) shows the ROC curve of the best model's test set performance from Tables II and III. For predicting

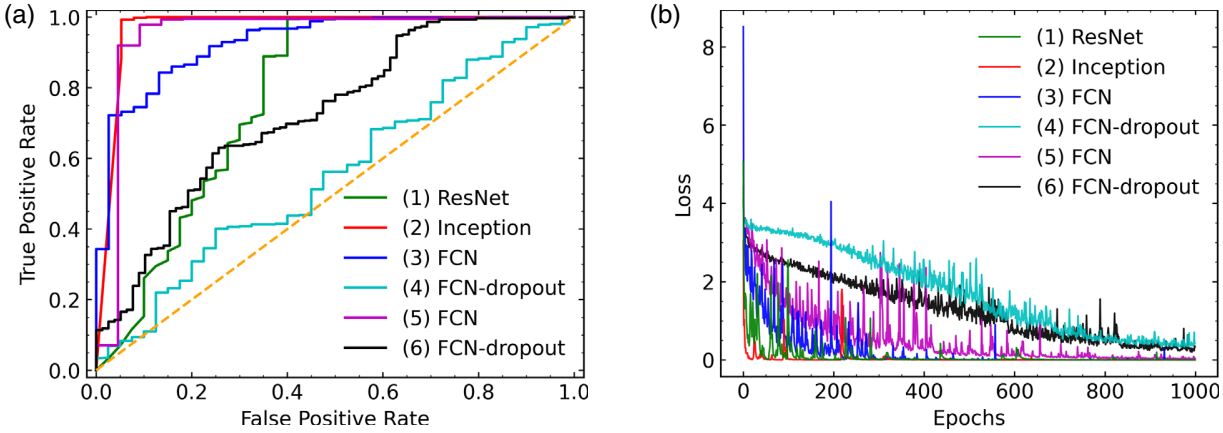


FIG. 14. Receiver operating characteristic (ROC) (a) and learning curve (b) of trend and event data modeling. For all prediction tasks (1–6) shown in the results in Table IV and Table V, the curves of the best model's test set is shown. The dashed orange line represents a random classifier in the ROC curve.

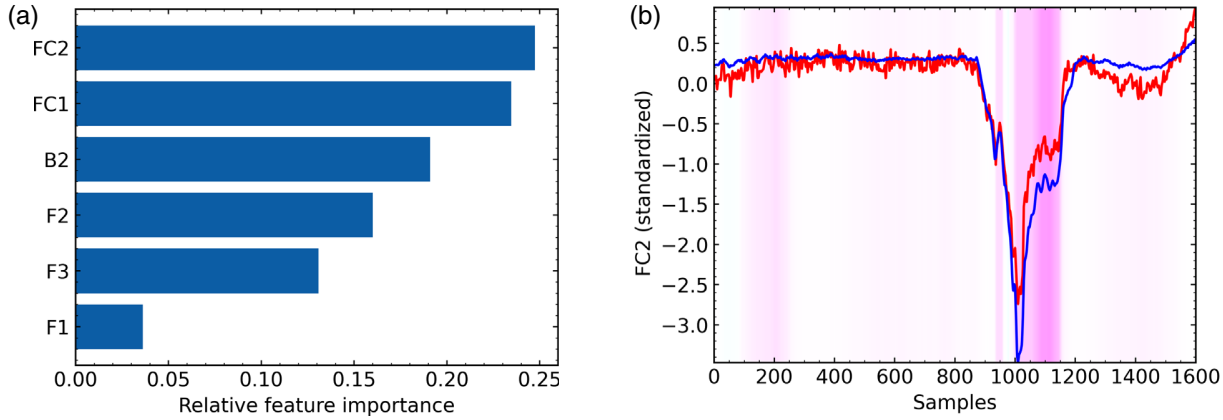


FIG. 15. Most important signals (a) and FC2 samples (b) for predicting follow-up breakdowns with event data. In addition to the most important samples (marked by the pink background), the average preceding signal for a subsequent healthy event (blue) and a breakdown event (red) is shown, respectively.

primary breakdowns, the FCN-dropout model (4, cyan) with $AR_t = 54.0\%$ is close to the orange dashed random classifier, where with $AR = 50.0\%$. Contrary, the FCN model (5, purple) for predicting follow-up breakdowns with $AR_t = 91.1\%$ covers a significantly larger area under the curve. The FCN-dropout model (6, black) combines the two curves, indicating that the predicted breakdowns were mostly follow-up breakdowns.

Similar to the trend data prediction, the threshold on $p(\hat{y}_{k+1} = 1|\boldsymbol{X}_k)$ can be selected. For example, there are two "edges" in the (5, purple) ROC curve at a false positive rate of about 0.05 and at 0.2. At the first "edge," ∼50% of all healthy events are classified correctly, and only 5% of breakdowns are falsely considered healthy. At the second "edge," ∼90% of all healthy events are classified correctly, but 20% of breakdowns are falsely classified as healthy. The selected threshold is dependent on the class weight, as we use $124,505 \times 2.5\% \approx 3113$ healthy and 479 breakdown events, and the effect on the machine availability of the application, as discussed in Sec. VI.

However, the number of epochs in our experiments is not fixed. The models are trained until the loss does not change significantly within 100 epochs, i.e., we use early stopping. Figure 14(b) shows the learning curve for the test set prediction of all the best models for 1000 epochs.

Models trained on trend data (1–3) converge faster than models trained on event data (4–6). In addition, models trained on follow-up breakdowns (2,5) converge faster than models trained on primary breakdowns (3,6). Also, the performance of classic machine learning models is closer to deep learning models in follow-up breakdowns compared to primary breakdowns. This indicates that correlations within the data and follow-up breakdowns are more linear compared to correlations within the data and primary breakdowns. The FCN-dropout model (4, cyan) for predicting primary breakdowns and the FCN-dropout model (5, black) fail to converge to a loss close to zero. This is in good agreement with the fact that those models achieve lower AR_t scores.

B. Explainable AI

Due to the poor performance for the prediction of primary breakdowns, only models for the prediction of follow-up breakdowns are considered for the explanation in this section.

The signals identified by the FCN as being most important for the prediction of follow-up breakdowns are shown in Fig. 15. The downstream Faraday cup signal (FC2) is classified as being most important (a) by the used models, but the difference to the other signals is not as significant as in Fig. 10. Further investigation showed that a specific portion of both Faraday cup signals, particularly the rising edge, was identified by the SHAP approach as being the most important region for breakdown prediction.

An example is shown with the downstream Faraday cup signals in Fig. 15(b). Here, the mean signal over all "healthy in the next pulse" events is plotted in blue and the mean over all "breakdown in the next pulse" events is plotted in red. The important samples in the signal, i.e., the SHAP values, are highlighted in pink. The most important area for the model is approximately 1000–1200 samples.

The reason for a relatively high noise in the red signal is twofold. First, there is higher variance in breakdown signals, as they generally vary in their shape. Second, follow-up breakdowns are generally lower in amplitude. This is due to the fact that after the machine is stopped as a consequence of a primary breakdown, its input power is gradually increased again to recover the nominal power. This leads to lower amplitudes in the follow-up breakdown signals. We mitigate this effect by standardizing each signal separately with its own mean and standard deviation. However, due to the lower amplitudes, the noise is more severe in follow-up breakdown signals. The increased deflection at the end of the red signal is also attributed

to this effect. Notably, our models do not focus on the noise or the deflection at the end, because the rising edge of both Faraday cup signals enables more general predictions.

The identified portion in the signal in Fig. 15 has been previously studied in detail [17,22]. The shape of the dark current signal is generally defined by several quantities. The fill time, i.e., the time for the rf pulse to propagate from the input to the output of the prototype CLIC structures, is generally in the order of 60 ns, which corresponds to 48 samples in the plot. As the rf pulse fills the structure of the individual cells, i.e., the subsection in the rf cavity, the cells begin to emit electrons. This results in a rising edge in the F1 signal which is comparable to the fill time of the structure. A similar transient behavior is then observed at the end of the rf pulse, as the structure empties and the cells stop emitting.

Breakdowns alter the surface of the rf cavity and thus change the emission properties of the structure. As a consequence, both the amplitude and shape of the signal are known to change markedly after breakdowns [73,75]. It is postulated that particular signal characteristics may then be associated with an increased probability of future breakdowns. Additionally, it has previously been proposed that fluctuations in the dark current signal may be associated with nascent breakdowns, however, these fluctuations have proven difficult to measure [22]. Such fluctuations constitute another phenomenon that could potentially be detected with the present framework. Notably, all previous observations seem compatible with the findings and explanations of our ML studies.

VI. FUTURE WORK

The goal of our study is twofold. First, we want to shed light on the physics associated with breakdowns through the insights gained with explainable AI. Second, we aim at supporting the development of data-driven algorithms for conditioning and operation of rf cavities based on machine learning. In this section, we further elaborate on these goals and future activities, starting from the results presented in the previous paragraphs.

A. Breakdown Physics

To further validate the explainable-AI findings in this work, future experiments will focus on the validation of the presence of a pressure rise prior to the occurrence of breakdowns, by using our simplified threshold-based model to provide an interlock signal. To make more insightful explanations, especially suited for the domain experts of CLIC, we will further improve the used explainable-AI algorithms. Current explainable-AI methods are developed and tested mostly with the goal to interpret images and highlight important areas for classification problems. Typical examples involve the recognition of characteristic features of animals, e.g., the ear of a cat. In images, those areas are self-explanatory and easy to understand by humans. However, explanations in time-series signals are harder to interpret (see Fig. 15). In the future, our work will focus on refining the model explanations by investigating the possibility of using understandable features and correlations to the important areas, e.g., the low mean value and high frequency in the important area of the red signal in Fig. 15. For this, we will build on existing work, which searches for correlations in the activations of the hidden CNN layers [61,76–79].

B. Model application

Investigations on the direct application of our models are ongoing. Here, the final model will be selected depending on the chosen task according to Tables II and III. For example, the FCN would be chosen for predicting follow-up breakdowns with event data, as it performs best. Below, we address several remaining challenges with which the model's performance could be improved and the potential of machine learning further exploited. Additionally, it is currently under evaluation of how the predictive methods can be embedded in the existing system by notifying an operator or by triggering an interlock before a predicted breakdown.

Model improvements.—To further advance the development of data-driven algorithms for conditioning and operation, we will test and improve our model with data from additional experiments. The accuracy of machine learning models is highly dependent on the quality of the data with which the model is trained. As such, the importance of continuous and consistent data logging during experiments is of primary importance during the study and further improvements are being discussed with the CLIC rf test stand team to (i) increase the logging frequency for both trend and event data, (ii) to implement signals of additional pressure sensitive sensors, e.g., vacuum gauges and vibration sensors, or (3) provide a means of accurate timing calibration in the test stand.

Model embedding.—As mentioned in Sec. II, it has previously been proposed that accelerating structures condition on the number of cumulative rf pulses and not solely on the cumulative number of breakdowns [25]. This also aligns with the intuition that conditioning is a process of material hardening caused by the stress of the applied electric field [26]. As such, possibilities are investigated to increase the applied field at a rate that still produces the hardening effect but refrains from inducing breakdowns unnecessarily frequently. Conversely, as conditioning typically requires on the order of hundreds of millions of pulses, it is highly desirable to minimize the number of pulses taken to reach high-field operation in order to reduce the electricity consumption and test duration. The optimal method may lie between these two scenarios, where our machine learning models come in to improve future conditioning algorithms.

Second, we focus on the possibility to derive operational algorithms that are planned to increase machine availability in modern high-gradient accelerators, exploiting our machine learning models. The basic idea is to maximize the availability of a future accelerator by dynamically detuning structures that are predicted to experience a breakdown, thus limiting the impact of breakdowns on the operation. The reduction in energy associated with doing so may then be compensated in one of two ways, either by powering an additional, spare structure in the beam line which is normally desynchronized, or alternatively, by temporarily increasing the voltage in the remaining structures until the arcing structure stabilizes again. In this scenario, the effect of false predictions of our model will directly affect the performance of the machine, and it is therefore of crucial importance to achieve sufficient accuracy in the predictions.

In a single rf structure, the approach discussed above is no longer valid. Currently, if a breakdown is detected, it is unclear if the breakdown is inevitable or if it may be avoided by taking an appropriate action. If the implemented response is one which interlocks the machine temporarily, a false prediction would then result in an unnecessary stop of the machine and hence a reduction in availability equal to that associated with the breakdown event. Thus, in such a scenario, a threshold on the probability of $p(\hat{y}_{k+1}|\boldsymbol{X}_k)$ is preferred such that the classification is healthy if the model is uncertain. Alternatively, a hybrid model [80] could be implemented, e.g., to enable machine operators to adjust the machine parameters if there are many predicted future breakdowns.

VII. CONCLUSION

In the work presented, a general introduction to data-driven machine learning models for breakdown prediction in rf cavities for accelerators was shown. Following the steps of transformation, exploration, modeling, and explanation, several state-of-the-art algorithms have been applied and have proven to be effective for our application. By interpreting the parameters of the developed models with explainable AI, we were able to obtain system-level knowledge, which we used to derive a fast, reliable, and threshold-based model.

We have shown that our models can predict primary breakdowns with 87.9% and follow-up breakdowns with an AR score of 98.7% using trend data. Thanks to the analyses carried out with explainable AI, we discovered that historical CLIC rf test bench data indicate that the pressure in the rf cavity begins to rise prior to the Faraday cup signals, in case of a breakdown. Our findings could enable the possibility to act before a breakdown is detected with the Faraday cup signal by setting a lower threshold on the vacuum signal. This would allow us to either avoid the breakdown development at an early stage or to take additional actions to preserve the beam quality.

Using event data, we achieved an AR score of 56.6% for predicting primary breakdowns and 89.7% on follow-up breakdowns, highlighting the low capabilities of the model to predict primary breakdowns but high performance on follow-up breakdowns. Focusing on the latter, explainable-AI points out that the last part of the rising edge in the Faraday cup signals has a high influence on the occurrence of breakdowns. Investigations to explain this behavior are currently ongoing but are supported by past studies on the subject.

Our code is publicly available[1] and provides a framework for the transformation, exploration, and modeling steps, which can be used to analyze breakdowns in other fields or domains.

[1] E. Sicking and R. Ström, From precision physics to the energy frontier with the Compact Linear Collider, Nat. Phys. **16**, 386 (2020).

[2] A. Grudiev and W. Wuensch, A new local field quantity describing the high gradient limit of accelerating structures, in *Proceedings of the 24th International Linac Conference, LINAC-2008, Victoria, BC, Canada* (2009).

[3] N. C. Lasheras, C. Eymin, G. McMonagle, S. Rey, I. Syratchev, B. Woolley, and W. Wuensch, Commissioning of XBox-3: A very high capacity X-band test stand, in *Proceedings of the 28th International Linac Conference* (2017), p. 4, https://cds.cern.ch/record/2304526/files/tuplr047.pdf.

[4] W. Wuensch, N. Catalán Lasheras, A. Degiovanni, S. Döbert, W. Farabolini, J. Kovermann, G. McMonagle, S. Rey, I. Syratchev, J. Tagg, L. Timeo, and B. Woolley, Experience operating an X-band high-power test stand at CERN, in *Proceedings of the 5th International Particle Accelerator Conference, IPAC-2014, Dresden, Germany, 2014* (JACoW, Geneva, Switzerland, 2014).

[5] A. Descoeudres, Y. Levinsen, S. Calatroni, M. Taborelli, and W. Wuensch, Investigation of the dc vacuum breakdown mechanism, Phys. Rev. ST Accel. Beams **12**, 092001 (2009).

[6] C. Adolphsen, Normal-conducting rf structure test facilities and results, in *Proceedings of the 20th Particle Accelerator Conference, PAC-2003, Portland, OR, 2003* (IEEE, New York, 2003), Vol. 1, pp. 668–672.

[7] Y. Donon, A. Kupriyanov, D. Kirsh, A. D. Meglio, R. Paringer, I. Rytsarev, P. Serafimovich, and S. Syomic, Extended anomaly detection and breakdown prediction in LINAC 4's rf power source output, in *Proceedings of International Conference on Information Technology and Nanotechnology, Samara, Russia* (IEEE, New York, 2020)

[8] E. Fol, R. Tomás, J. Coello De Portugal, and G. Franchetti, Detection of faulty beam position monitors using unsupervised learning, Phys. Rev. Accel. Beams **23**, 102805 (2020).

[9] C. Emma, A. Edelen, M. J. Hogan, B. O'Shea, G. White, and V. Yakimenko, Machine learning-based longitudinal

phase space prediction of particle accelerators, Phys. Rev. Accel. Beams **21**, 112802 (2018).

[10] O. Convery, L. Smith, Y. Gal, and A. Hanuka, Uncertainty quantification for virtual diagnostic of particle accelerators, Phys. Rev. Accel. Beams **24**, 074602 (2021).

[11] S. R. Xie, G. R. Stewart, J. J. Hamlin, P. J. Hirschfeld, and R. G. Hennig, Functional form of the superconducting critical temperature from machine learning, Phys. Rev. B **100**, 174513 (2019).

[12] C. Tennant, A. Carpenter, T. Powers, A. Shabalina Solopova, L. Vidyaratne, and K. Iftekharuddin, Superconducting radio-frequency cavity fault classification using machine learning at Jefferson Laboratory, Phys. Rev. Accel. Beams **23**, 114601 (2020).

[13] B. J. Woolley, High power X-band rf test stand development and high power testing of the CLIC Crab Cavity, Ph.D. thesis, Lancaster University, 2015.

[14] Z. D. Farkas, H. A. Hoag, G. A. Loew, and P. B. Wilson, SLED: A method of doubling SLAC's energy, in *Proceedings of the 9th International Conference on High-Energy Accelerators, HEACC-1974, Stanford, CA, 1974* (1974), https://s3.cern.ch/inspire-prod-files-0/068ed22dae5ba0bab8bcf19f9dcac092.

[15] B. Woolley, I. Syratchev, and A. Dexter, Control and performance improvements of a pulse compressor in use for testing accelerating structures at high power, Phys. Rev. Accel. Beams **20**, 101001 (2017).

[16] R. Zennaro *et al.*, High power tests of a prototype X-band accelerating structure for CLIC, in *Proceedings of 8th International Particle Accelerator Conference, IPAC-2017, Copenhagen, Denmark, 2017* (JACoW, Geneva, Switzerland, 2017), pp. 4318–4320.

[17] J. Paszkiewicz, P. Burrows, and W. Wuensch, Spatially resolved dark current in high gradient traveling wave structures, in *Proceedings of the 10th International Particle Accelerator Conference, Melbourne, Australia, 2019* (JACoW, Geneva, Switzerland, 2019).

[18] T. G. Lucas, T. Argyopolous, M. J. Boland, N. Catalan-Lasheras, R. P. Rassool, C. Serpico, M. Volpi, and W. Wuensch, Dependency of the capture of field emitted electron on the phase velocity of a high-frequency accelerating structure, Nucl. Instrum. Methods Phys. Res., Sect. A **914**, 46 (2019).

[19] D. Banon Caballero *et al.*, Dark current analysis at CERN's X-band facility, in *Proceedings of the 10th International Particle Accelerator Conference, Melbourne, Australia* (JACoW, Geneva, Switzerland, 2019).

[20] B. Woolley, I. Syratchev, and A. Dexter, Control and performance improvements of a pulse compressor in use for testing accelerating structures at high power, Phys. Rev. Accel. Beams **20**, 101001 (2017).

[21] B. Woolley, G. Burt, A. C. Dexter, R. Peacock, W. L. Millar, N. Catalan Lasheras, A. Degiovanni, A. Grudiev, G. Mcmonagle, I. Syratchev, W. Wuensch, E. Rodriguez Castro, and J. Giner Navarro, High-gradient behavior of a dipole-mode rf structure, Phys. Rev. Accel. Beams **23**, 122002 (2020).

[22] J. Paszkiewicz, Studies of breakdown and pre-breakdown phenomena in high-gradient accelerating structures, Ph.D. thesis, St. John's College, Oxford, 2021.

[23] J. W. Kovermann, Comparative studies of high-gradient rf and dc breakdowns, Ph.D. thesis, CERN, 2010.

[24] H. Sak, A. Senior, and F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in *Proceedings of the Annual Conference of the International Speech Communication Association* (2014), https://static.googleusercontent.com/media/research.google.com/de//pubs/archive/43905.pdf.

[25] J. Giner Navarro, Breakdown studies for high-gradient rf warm technology in: CLIC and hadron therapy linacs, Ph.D. thesis, University of Valencia (main), 2016.

[26] A. Korsbäck, F. Djurabekova, L. M. Morales, I. Profatilova, E. R. Castro, W. Wuensch, S. Calatroni, and T. Ahlgren, Vacuum electrical breakdown conditioning study in a parallel plate electrode pulsed dc system, Phys. Rev. Accel. Beams **23**, 033102 (2020).

[27] V. A. Dolgashev and S. G. Tantawi, Study of basic breakdown phenomena in high gradient vacuum structures, in *Proceedings of the 25th International Linear Accelerator Conference, LINAC-2010, Tsukuba, Japan* (KEK, Tsukuba, Japan, 2010), Vol. 2.

[28] P. Craievich *et al.*, Consolidation and extension of the high-gradient LINAC rf Technology at PSI, in *Proceedings of the 29th International Linear Accelerator Conference, LINAC-2018, Beijing, China* (JACoW, Geneva, Switzerland, 2018), THPO115.

[29] L. Felsberger, A. Apollonio, T. Cartier-Michaud, A. Müller, B. Todd, and D. Kranzlmüller, Explainable deep learning for fault prognostics in complex systems: A particle accelerator use-case, in *Machine Learning and Knowledge Extraction. CD MAKE 2020*, Lecture Notes in Computer Science Vol. 12279 (Springer, Cham, 2020).

[30] J. Guo, Z. Li, and M. Li, A review on prognostics methods for engineering systems, IEEE Trans. Reliab. **69**, 1110 (2019).

[31] Y. Ran, X. Zhou, P. Lin, Y. Wen, and R. Deng, A survey of predictive maintenance: Systems, purposes and approaches, arXiv:1912.07383.

[32] A. Calaprice, *The Ultimate Quotable Einstein* (Princeton University Press, Princeton, NJ, 2011).

[33] N. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. **46**, 175 (1992).

[34] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).

[35] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn. **20**, 273 (1995).

[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT press, Cambridge, MA, 2016).

[37] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, Deep learning for time series classification: A review, Data Min. Knowl. Discovery **33**, 917 (2019).

[38] M. Christ, A. W. Kempa-Liehr, and M. Feindt, Distributed and parallel time series feature extraction for industrial big data applications, CoRR **abs/1610.07717** (2016), http://arxiv.org/abs/1610.07717.

[39] S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Curran Associates Inc., Red Hook, 2017), https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[40] S. Wold, K. Esbensen, and P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. **2**, 37 (1987)
[41] G. Hinton and S. Roweis, Stochastic neighbor embedding, in *Proceedings of the Advances in Neural Information Processing Systems, NIPS 2002* (2002), Vol. 15, pp. 857–864, https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.
[42] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).
[43] J.-Y. Franceschi, A. Dieuleveut, and M. Jaggi, Unsupervised scalable representation learning for multivariate time series, in *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2019* (2019), Vol. 32, pp. 4650–4661, https://papers.nips.cc/paper/2019/file/53c6de78244e9f528eb3e1cda69699bb-Paper.pdf.
[44] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006), p. 738.
[45] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proceedings of the Advances in Neural Information Processing Systems, NIPS 2012* (2012), Vol. 25, https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
[47] R. Collobert and J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th International Conference on Machine Learning, ICML'08* (2008), http://machinelearning.org/archive/icml2008/papers/391.pdf.
[48] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, Convolutional neural networks for time series classification, J. Syst. Eng. Electron. **28**, 162 (2017).
[49] Z. Wang, W. Yan, and T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, in *Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, 2017* (IEEE, New York, 2017).
[50] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv:1502.03167.
[51] C. Szegedy, S. Ioffe, and V. Vanhoucke, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, CoRR **abs/1602.07261** (2016), http://arxiv.org/abs/1602.07261.
[52] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P. A. Muller, and F. Petitjean, InceptionTime: Finding AlexNet for time series classification, Data Min. Knowl. Discovery **34**, 1936 (2020).
[53] M. Abadi *et al.*, TensorFlow: A system for large-scale machine learning, in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA* (2016), https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.
[54] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. **27**, 861 (2006).
[55] S. Reardon, Rise of robot radiologists, Nature (London) **576**, S54 (2019).
[56] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, Can machine-learning improve cardiovascular risk prediction using routine clinical data?, PLoS One **12**, e0174944 (2017).
[57] H. Lakkaraju and C. Rudin, Learning cost-effective and interpretable treatment regimes, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (PMLR, New York, 2017), pp. 166–175.
[58] M. A. Qureshi and D. Greene, Eve: Explainable vector based embedding technique using wikipedia, J.Intell. Inf. Syst. **53**, 137 (2019).
[59] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, Explainable AI in industry (2020).
[60] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, Not just a black box: Interpretable deep learning by propagating activation differences, CoRR **abs/1605.01713** (2016), arXiv:1605.01713.
[61] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, How can I explain this to you? An empirical study of deep neural network explanation methods, in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20* (2020), Vol. 33, p. 12.
[62] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One **10**, e0130140 (2015).
[63] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the predictions of any classifier, arXiv:1602.04938.
[64] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, New York, 2018), pp. 839–847.
[65] K. Simonyan, A. Vedaldi, and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv:1312.6034.
[66] L. Song, A. J. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo, Supervised feature selection via dependence estimation, CoRR **abs/0704.2668** (2007), http://arxiv.org/abs/0704.2668.
[67] I. Guyon and A. Elisseeff, An introduction to variable
and feature selection, J. Mach. Learn. Res. **3**, 1157 (2003).
[68] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. **46**, 389 (2002).
[69] M. A. Wojtas and Ke Chen, Feature importance ranking for deep learning, in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (2020), Vol. 33, https://proceedings.neurips.cc/paper/2020/file/36ac8e558ac7690b6f44e2cb5ef93322-Paper.pdf.
[70] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, Kernel feature selection via conditional covariance minimization, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., New York, 2017), Vol. 30.

[71] L. S. Shapley, *A Value for n-Person Games* (Princeton University Press, Princeton, NJ, 2016).

[72] P. Simard, D. Steinkraus, J. Platt *et al.*, Best practices for convolutional neural networks applied to visual document analysis, in *Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, United Kingdom, 2003* (IEEE, New York, 2003), Vol. 3.

[73] E. Z. Engelberg, J. Paszkiewicz, R. Peacock, S. Lachmann, Y. Ashkenazy, and W. Wuensch, Dark current spikes as an indicator of mobile dislocation dynamics under intense dc electric fields, Phys. Rev. Accel. Beams **23**, 123501 (2020).

[74] E. Z. Engelberg, Y. Ashkenazy, and M. Assaf, Stochastic Model of Breakdown Nucleation under Intense Electric Fields, Phys. Rev. Lett. **120**, 124801 (2018).

[75] D. Banon-Caballero, M. Boronat, N. Catalán Lasheras, A. Faus-Golfe, B. Gimeno, T. G. Lucas, W. L. Millar, J. Paszkiewicz, S. Pitman, V. Sánchez Sebastián, A. Vnuchenko, M. Volpi, M. Widorski, W. Wuensch, and V. del Pozo Romano, Dark current analysis at CERN's X-band Facility, in *Proceedings of the 10th International Particle Accelerator Conference, Melbourne, Australia, 2019*, edited by M. Boland, H. Tanaka, D. Button, R. Dowd, V. R. W. Schaa, and E. Tan (JACoW Publishing, Geneva, Switzerland, 2019), pp. 2944–2947.

[76] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, Dynamical variational autoencoders: A comprehensive review, CoRR **abs/2008.12595** (2020), https://arxiv.org/abs/2008.12595.

[77] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in *Proceedings of the International Conference on Machine Learning* (PMLR, New York, 2018), pp. 2668–2677.

[78] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, Towards automatic concept-based explanations, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' AlchéBuc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, New York, 2019), Vol. 32.

[79] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, On completeness-aware concept-based explanations in deep neural networks, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, New York, 2020), Vol. 33, pp. 20554–20565.

[80] C. Obermair, M. Maciejewski, F. Pernkopf, Z. Charifoulline, A. Apollonio, and A. Verweij, Machine learning with a hybrid model for monitoring of the protection systems of the LHC, in *Proceedings of 12th International Particle Accelerator Conference, IPAC-2021, Campinas, SP, Brazil* (JACoW, Geneva, Switzerland, 2021).

Paper 5

Example or Prototype? Learning Concept-Based Explanations in Time-Series

Example or Prototype? Learning Concept-Based Explanations in Time-Series

Christoph Obermair OBERMAIR@TUGRAZ.AT
Graz University of Technology, Graz, Austria
CERN, Geneva, Switzerland
Alexander Fuchs FUCHS@TUGRAZ.AT
Franz Pernkopf PERNKOPF@TUGRAZ.AT
Graz University of Technology, Graz, Austria
Lukas Felsberger LUKAS.FELSBERGER@CERN.CH
Andrea Apollonio ANDREA.APOLLONIO@CERN.CH
Daniel Wollmann DANIEL.WOLLMANN@CERN.CH
CERN, Geneva, Switzerland

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

With the continuous increase of deep learning applications in safety critical systems, the need for an interpretable decision-making process has become a priority within the research community. While there are many existing explainable artificial intelligence algorithms, a systematic assessment of the suitability of global explanation methods for different applications is not available. In this paper, we respond to this demand by systematically comparing two existing global concept-based explanation methods with our proposed global, model-agnostic concept-based explanation method for time-series data. This method is based on an autoencoder structure and derives abstract global explanations called "prototypes". The results of a human user study and a quantitative analysis show a superior performance of the proposed method, but also highlight the necessity of tailoring explanation methods to the target audience of machine learning models.

Keywords: Explainable AI, Concept Explanations, Time-Series.

1. Introduction

Deep learning methods have conquered nearly every aspect of machine learning applications due to their flexibility and predictive power. However, they did not yet gain the same interest in safety-critical applications, due to their "black box" behavior (Beítez et al. (1997)). Especially in safety critical applications, wrong decisions can have severe impact on human health, *e.g.* in medical diagnosis (Reardon (2019); Weng et al. (2017)) or financial assets and reputation of large scale projects such as particle accelerators (Obermair et al. (2022)). In these cases, experts cannot simply rely on automatically generated predictions and are often legally obliged to state reasons for their decisions (Goodman and Flaxman (2017)). Therefore, the demand for methods that allow for the interpretation of black box models has been increasing and a wide variety of eXplainable Artificial Intelligence (XAI) algorithms were proposed in recent years.

The currently most popular XAI algorithms (Ribeiro et al. (2016); Bach et al. (2015); Shrikumar et al. (2016); Lundberg and Lee (2017); Simonyan et al. (2013); Chattopadhay et al. (2018)) are *relevance-based* methods capable of highlighting the parts of the data which are important for model predictions. Considering a handwritten digit classification problem using the MNIST dataset (Deng (2012)), for example, highlighting relevant pixels representing a particular number is an intuitive interpretation for humans.

Concept-based explanations (Kim et al. (2018); Yeh et al. (2020)) represent an alternative to highlighting important parts of the data. While there exist multiple definitions of *concepts* across the literature, we define a concept as *explanatory data containing all relevant properties that allow humans to make the same decisions as the black box model.* Typically, concepts are provided by (1) data examples, *i.e. explanation-by-example*, or (2) artificial data containing the most relevant information, *i.e. prototypes.* In the example of handwritten digit classification, showing the image of a typical digit 'one' from the available data would be an explanation-by-example, while showing an artificially created example of the digit 'one' with its main properties, *e.g.* the straight vertical line, would be a prototype.

In a recent empirical study conducted within a group of non-machine-learning experts, (Jeyakumar et al. (2020)) showed superior performance of a concept-based explanation method compared to relevance-based methods for time-series data. Explaining the non-intuitive nature of time-series data to non-machine-learning experts is a common task in safety critical applications, *e.g.* when explaining heart beat signals to medical professionals and patients. Consequently, concept-based explanations are an important tool in this domain. However, explanation-by-example and prototypes have not been compared in detail yet, although they belong to the main types of existing concept-based explanation methods.

Contribution. In this work, we investigate the advantages and disadvantages of explanation-by-example or prototypes for time-series explanations, depending on whether the target audience is users or model developers. Initially, we define a concept mathematically and denote concept properties to increase the explanation confidence. Consequently, we propose a model-agnostic concept-based XAI method[1], relying on an autoencoder using prototypes. We then compare our model-agnostic prototype (MAP) method to an explanation-by-example (EBE) (Jeyakumar et al. (2020)) and a model-specific prototype (MSP) (Gee et al. (2019)) explanation method with a human user study and a quantitative analysis.

Human User Study Details. For the conducted human user study, we utilized the ECG200 (Olszewski (2001)) dataset containing heartbeat signals and an artificial dataset reproducing signals from machine sensors in a noisy environment. Participants were asked to classify the time-series signals from the dataset, using the concept explanations which we provided. In total, 75 participants classified 3480 time-series signals based on explanation-by-example or prototypes derived from the different methods. The survey shows that our method is preferred, but also highlights the importance to distinguish between target audiences when comparing XAI methods.

Paper Structure. We first give an overview of related XAI work, followed by a formal definition of a concept and its properties. We then introduce our XAI method and our

1. https://github.com/cobermai/concep_based_explanations

study details. Finally, we discuss the results, and present future work in the domain of particle accelerators.

2. Related Work

In this section, we highlight the need for concept explanations, which are model-agnostic, applicable to time-series data, and tested and optimized for their target audience. With the increasing amount of time-series data available, hundreds of time-series classification methods have been recently proposed. Different methods are frequently based on nearest neighbors (Bagnall et al. (2017)), ensemble classifiers (Lines et al. (2018)), or convolutional neural networks (Fawaz et al. (2019)).

Many of the recently proposed XAI methods target the interpretation of such time-series classification methods (Rojat et al. (2021)). This is especially relevant for safety critical applications, where time-series data is a common data format. Tjoa and Guan (2020) provide a list of different XAI methods for medical applications as an example for safety critical applications. A recent summary from AlRegib and Prabhushankar (2022) highlights the small amount of model-agnostic XAI approaches and underlines the importance of human evaluation of such approaches. Amazons Mechanical Turk enables a relatively fast way to derive human non-expert evaluations without a bias, and is commonly used in XAI studies as in Jeyakumar et al. (2020), Lundberg and Lee (2017), Ribeiro et al. (2016), and Kim et al. (2018). It is more difficult to choose an intentional bias. For example, a bias towards the characteristics of the research community in safety-critical applications.

In the following subsections, we provide an overview of relevant concept explanation methods, distinguishing between methods using explanation-by-example and methods using prototypes to visualize their concepts. For each method, we emphasize whether the model is model-specific and whether the explanations are local or global. Local explanations, analyze the black-box predictions of each data sample, *i.e.* an *instance*, separately, while global explanations investigate all predictions at once.

2.1. Concept visualization with explanation-by-example

Kim et al. (2018); Yeh et al. (2020) describe concepts as a set of implicit vectors. To visualize a concept, the instance/example closest to the vector is extracted from the model specific architecture. Jeyakumar et al. (2020) cluster instances with similar activations in the last layer of a deep neural network. They use the cosine similarity as a similarity measure. The access of the activations makes this method model-specific. Explanation-by-example is frequently extended to show only relevant segments of examples. Chen et al. (2019); Das et al. (2020) show image patches like the ear of cat as examples. Guidotti et al. (2020) propose a model-agnosic method, which generates relevant example segments, using decision trees. These segments are frequently called shapelets. In Mochaourab et al. (2022) global explanations are derived from relevance-based explanation using Sobol's indices, *i.e.* a variance-based sensitivity analysis.

2.2. Concept visualization with prototypes

Prototype based methods aim at defining representative concept prototypes for model explanation (Bien and Tibshirani (2011)). Li et al. (2017); Gee et al. (2019) train prototypes with an autoencoder. A classifier is trained in parallel. This classifier utilizes the euclidean distance of the prototypes and the latent space of the autoencoder as an input. Here, classification and explanation are combined in the same model, which makes it model specific. In a similar way Zhang et al. (2020) derive one prototype per class with a model specific attention prototype network. Tang et al. (2020) generate time-series shapelets by combining concept-based and relevance-based methods.

The presented list of state-of-the-art methods, highlights the frequent use of prototypes and examples for visualization. For these methods, it has not been evaluated, which visualization technique is best in helping humans to reach similar accuracy as the black box model. This topic will be mathematically approached in the next section.

3. Concept Definition and Properties

Consider a training set of N instances $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, where each instance $\mathbf{x}_n \in \mathbb{R}^p$ has a corresponding label $y_n \in \mathbb{N}^t$, and a black box model $f(\cdot)$, *e.g.* a pretrained deep neural network, which approximates these labels $\hat{y}_n = f(\mathbf{x}_n)$. An explainer model is then used to derive a set of M concept explanations $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_M\}$ with predictions $\hat{\mathbf{Y}} = \{f(\hat{\mathbf{x}}_1), ..., f(\hat{\mathbf{x}}_M)\}$, where each explanation $m = 1, ..., M$ corresponds to a reconstructed concept $\hat{\mathbf{x}}_m$.

XAI methods are often evaluated through dedicated questionnaires (Holzinger et al. (2020)), asking its users to state their subjective assessment of the given explanation. To provide an objective evaluation of XAI methods, a human perceiver $s(\cdot)$ of an explanation should be able to find the correct label for unseen instances on their own. Showing all concept explanations $\hat{\mathbf{X}}$ and their corresponding labels $\hat{\mathbf{Y}}$ to users in the target audience, the *concept receptivity* is measured by the accuracy of the users when labeling new instances $\mathbf{x}_n$.

Definition 1 (Concept Receptivity) *A human perceiver $s(\cdot)$ has a concept receptivity r, which is the ability to find the label $\hat{y}_n$ for random instances $\mathbf{x}_n$ given the reconstructed concepts $\hat{\mathbf{X}}$ with labels $\hat{\mathbf{Y}}$*

$$r(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\hat{y}_n = s(\mathbf{x}_n, \hat{\mathbf{X}}, \hat{\mathbf{Y}})}, \tag{1}$$

where $\mathbb{1}$ is an indicator function.

Human evaluations of explanations are labor-intensive. To this end, we further propose a quantitative evaluation method. Explainer models frequently use a transformation function to derive lower dimensional features $\mathbf{z}_i = g(\mathbf{x}_i)$, where $\mathbf{z}_i \in \mathbb{R}^q, \mathbf{x}_i \in \mathbb{R}^p$, and $q < p$. In order for this latent space to faithfully represent the input space, the relation of instances in the latent space, should be similar to the instances in the input space.

Definition 2 (Representability) *We specify the similarity of two instances* $\mathbf{x}_j$ *and* $\mathbf{x}_n$ *with the conditional probability*

$$p(\mathbf{x}_j|\mathbf{x}_n) = \frac{\exp(-||\mathbf{x}_n - \mathbf{x}_j||^2/2\sigma_n^2)}{\sum_{k\neq n}\exp(-||\mathbf{x}_n - \mathbf{x}_k||^2\sigma_n^2)}, \tag{2}$$

assuming a Gaussian distribution of data points with standard deviation σ *(Van der Maaten and Hinton (2008)). With this definition, we compare the conditional probabilities* $P_n = p(\mathbf{x}_j|\mathbf{x}_n)$ *and* $Q_n = p(\mathbf{z}_j|\mathbf{z}_n)$*, between input instances* $\mathbf{x}_j$ *and their latent space activations* $\mathbf{z}_j$*. We, therefore, determine the Kullback-Leibler (KL) divergence between conditional probabilities of one instance* n*, to all other* N *instances in the dataset. Notably,* $KL_{P_n=Q_n}(P_n||Q_n) = 0$ *indicates that distribution* P_n *equals* Q_n*. The sum of all KL divergences is the concept representability*

$$\phi_c = \sum_n^N KL(p(\mathbf{x}_j|\mathbf{x}_n) \parallel p(\mathbf{z}_j|\mathbf{z}_n)) = \sum_n^N \sum_j^N p(\mathbf{x}_j|\mathbf{x}_n) \log \frac{p(\mathbf{x}_j|\mathbf{x}_n)}{p(\mathbf{z}_j|\mathbf{z}_n)}. \tag{3}$$

Similarly, we determine how well the reconstructed concepts represent the input. In order to make the M *concepts comparable to the* N *input instances, we look for the nearest concept of each input instance in the latent space in terms of the* L_2*-Norm,* $\arg\min_{\hat{\mathbf{x}}_m} ||g(\mathbf{x}_n) - g(\hat{\mathbf{x}}_m)||_2$*. Hence, we obtain the reconstructed concepts in the input space,* $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_N\}$*. We define the reconstructed concept representability as the sum of all KL divergences,*

$$\phi_{cr} = \sum_n^N KL(p(\mathbf{x}_j|\mathbf{x}_n) \parallel p(\hat{\mathbf{x}}_j|\hat{\mathbf{x}}_n)). \tag{4}$$

Fig. 1 depicts the concept representability ϕ_c and the reconstructed concept representability ϕ_{cr}. The input $\mathbf{x}_n$ consist of two blue signals of class one and one red signal of class two. Two concepts $\mathbf{c}_1, \mathbf{c}_2$ are derived from the latent space $\mathbf{z}_n$ with k-means. The red signal is reconstructed with $\mathbf{c}_2$. The two blue signals are closest to $\mathbf{c}_1$, and their reconstructed concept is therefore equal. Hence, the similarity of the input signals is well reflected by the concepts.

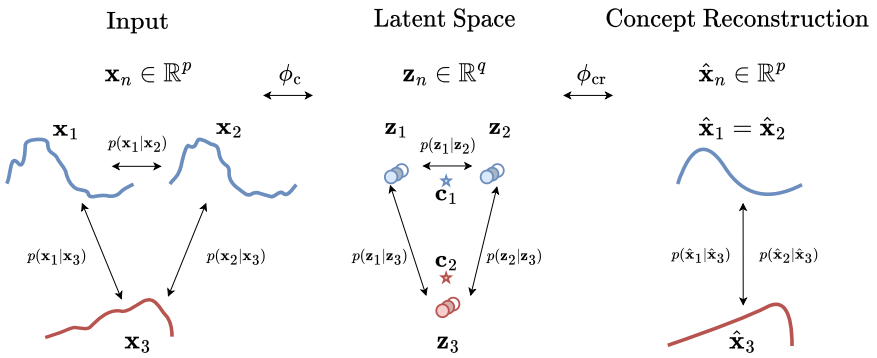


Figure 1: Example with three instances of two classes in red and blue. Two concepts have been reconstructed from the latent space.

4. Model-Agnostic Prototype Method

Our proposed method uses an autoencoder architecture, shown in Fig. 2, consisting of an encoder function $g(\cdot)$ that maps each instance n onto a latent space $\mathbf{z}_n \in \mathbb{R}^q$, and a decoder function $h(\cdot)$ that transforms the latent space back to the original input space $\mathbb{R}^p$. Using the latent space of the training set, we infer the concepts $\mathbf{C} = \{\mathbf{c}_1, ..., \mathbf{c}_M\}$ with k-means ($k = M$), where concepts are considered to be prototypes. A prototype enables the inference of M reconstructed concepts through $\hat{\mathbf{x}}_m = h(\mathbf{c}_m)$, using $\mathbf{c}_m$ within the latent space of the model.

Our method is trained independently of the black box model and is therefore model-agnostic. This not only enables to use any existing model without modifications, it also enables to derive explanations for already trained models. Furthermore, we argue that model-specific explanation methods, that access the activations of a hidden layer from a trained black box model, infer worse reconstructions, as detailed information necessary for the reconstruction is lost in the process of optimizing the weights for classification. Unlike other autoencoder methods (Gee et al. (2019); Li et al. (2017)), we derive our concepts directly from the latent space, *i.e.* the activations of the last encoder layer, instead of optimizing the concepts during training. We also employ a similarity loss for the latent space to diversify the concepts in Eq.7. In practice, this leads to more robust training, faster convergence, and more meaningful concepts.

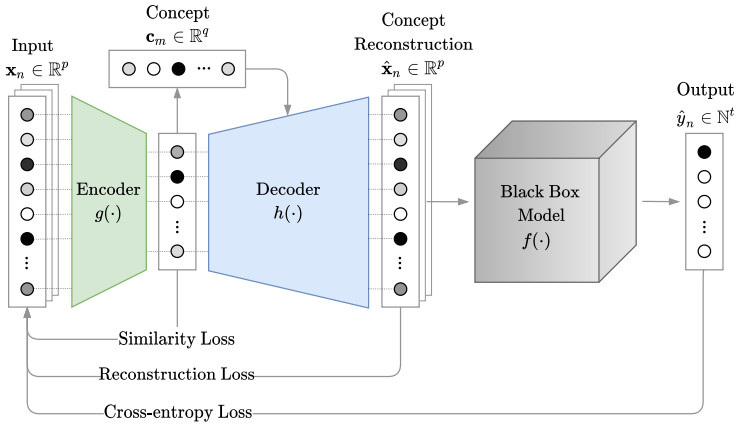


Figure 2: Model architecture used for the MAP explainer. Given a trained black box model, we fit an autoencoder to reconstruct the input data and to recreate the output of the black box model. The concepts are derived from the autoencoder latent space and are optimized to be diverse.

During optimization of the autoencoder weights, we maximize the ability to reconstruct both the input, and the exact prediction in terms of softmax outputs of the black box model. To regularize the concepts, we also employ a similarity loss during training. For the reconstruction loss, we use the mean-squared-error,

$$R(g, h, \mathbf{X}) = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}_n - h\left(g(\mathbf{x}_n)\right)\right)^2 . \tag{5}$$

For classification tasks the ability to reconstruct the model prediction, is measured via the categorical-cross-entropy loss,

$$C(g,h,f,\mathbf{X}) = -\sum_{n=1}^{N}\left(\arg\max_{\hat{y}} f(\mathbf{x}_n)\right)\log f\left(h\left(g(\mathbf{x}_n)\right)\right). \tag{6}$$

Diverse concepts are obtained by penalizing non-orthogonality between two different concepts $\mathbf{c}_i, \mathbf{c}_j \in \mathbf{C}$. Specifically, we define the similarity loss as the scaled sum of their inner products (Yeh et al. (2020))

$$S(\mathbf{C}) = \frac{\sum_{i\neq j}\mathbf{c}_i^T\mathbf{c}_j}{M(M-1)}, \tag{7}$$

where the concepts are the cluster centers of the latent space, derived with k-means[2]. The complete learning objective is given as follows. Notably y_n is not required, which enables an unsupervised training of model-agnostic prototypes,

$$\mathcal{L}(g,h,\mathbf{X}) = R(g,h,\mathbf{X}) + \lambda_C C(g,h,f,\mathbf{X}) + \lambda_S S(\mathbf{C}). \tag{8}$$

4.1. Model Structure

We use two different autoencoder structures (for full details, see appended code[1]), based on an extensive sensitivity analysis, and relevant literature from Agarap (2018) and O'Shea and Nash (2015). In this section, we first validate our model and show the effect of the hyperparameters λ_C and λ_S in Section 4.2. For this task we use a three layer convolutional autoencoder architecture with a 3x3 kernel and a filter size of 32, 64, and 1 for both the encoder and the decoder. Additionally, both the encoder and decoder, use ReLU activations in the first two layers, and a sigmoid activation in the last layer.

For the explanation of time-series classification, in Section 5 and 6, we use a one layer encoder, L1-activity regularization and a normalized output. This means that our method finds a linear mapping of the input signal to the concepts. Multivariate time-series are flattened before the encoder. We set the hyperparameters to $\lambda_C = 1$ and $\lambda_S = 1$, and monitor that all loss terms converge. Furthermore, we used a three layer neural network, with 300 neurons per layer and a sigmoid activation in the second layer as a decoder.

For both autoencoders, we set the latent space size to five times the number of concepts and use the ADAM optimizer. This enables all loss terms to converge, while keeping the latent space small.

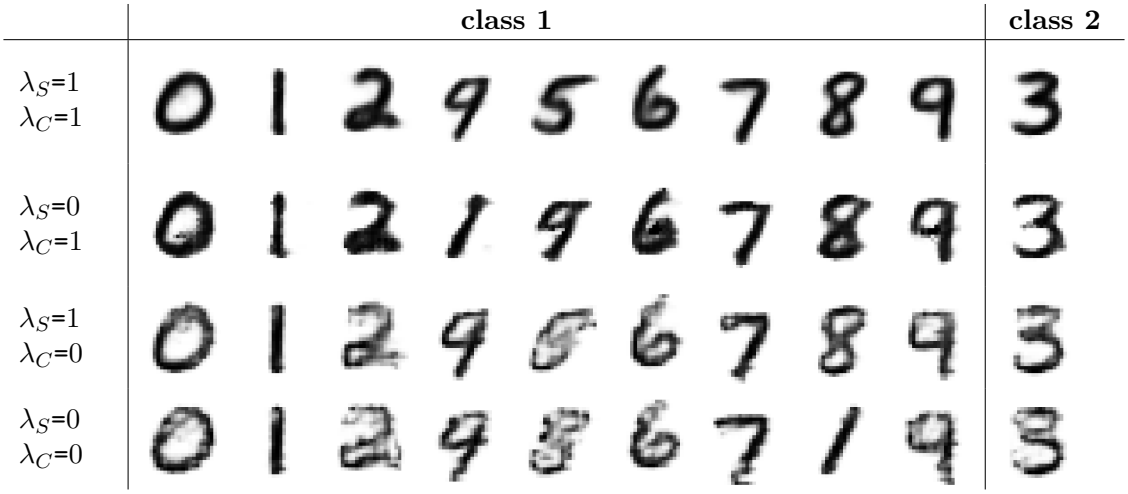
4.2. Model Validation

We validate our method by explaining a classifier, trained to predict whether an instance of the MNIST dataset of handwritten digits (Deng (2012)) contains the digit 'three'. As a classifier, we use a four layer neural network. Specifically, it consists of two convolutional layers of size 32 and 64, followed by two fully connected layers of size 128 and 10. All layers but the last use ReLU activations, where a softmax activation is used. In addition, we use max pooling after the second layer, and 0.2 and 0.4 dropout after the last two layers. The classifier is trained with the ADAM optimizer, achieving an accuracy of 99.8%.

2. In practice, this means that k-means is applied on each batch during training.

Ten model-agnostic prototypes ($M = 10$) were reconstructed, shown in Fig. 3, with different hyperparameters λ_C and λ_S. These prototypes, were manually sorted. The prototypes, calculated with $\lambda_C = 1$ and $\lambda_S = 1$, manage to represent all digits in the dataset, except for an overlap of digits 'four' and 'nine'. Without similarity loss ($\lambda_S = 0$), there are two concepts for the digit 'one', an overlap in the digits 'five' and 'nine', and digit 'four' is missing. Without classification loss ($\lambda_C = 0$), all numbers are represented with lower reconstruction performance compared to the first row. Finally, the effect of setting both hyperparameters to 0 is shown ($\lambda_C = 0$ and $\lambda_S = 0$). This means the reconstructions are blurry, the number 'eight' is missing, and the number 'one' occurs twice. With this simple example, we demonstrate the effectiveness of our model on an easily interpretable and well known dataset. The same approach will be applied to time-series classification problems, which are harder to interpret for humans.

Figure 3: Validation of our MAP method using MNIST with and without similarity and classification losses, with M=10. The black box model was trained to classify whether an image contains the digit 'three'.



5. Methodology

5.1. Modeling Methodology

We conduct a quantitative analysis on 12 datasets and a human user study, where we assess two of the datasets qualitatively. The selection of the datasets is based on signals frequently used in safety critical applications. We derive 11 of the datasets from the UCR archive Dau et al. (2019) and create one artificial dataset on our own. We address details for the human user study and the experimental methods in Section 5.2. The results of the quantitative analysis and a human user study, are reported in Section 6.1 and Section 6.2, respectively.

5.2. Human User Study[3]

We analyze the suitability of the three different methods in the context of safety critical applications empirically, with one survey[1] per method. For each survey, participants labeled 15 instances from two different datasets. These 15 instances were drawn randomly from the dataset to ensure fair comparison. To choose the right label, participants were given the concept explanations for each class in the dataset. Out of all correct answers, we then calculated the concept receptivity, as described in Eq. 1.

Participants. Our study was distributed within our research community, collecting a total of 3480 answers from 75 students and research staff working in the field of safety critical applications. The bias, introduced due to sampling in this research community, is intentional in order to optimize the models towards their future users. In the beginning of the survey, participants were asked to indicate whether they have prior knowledge in the field of machine learning. People who answered positively to this question were classified as typical *developers* of ML methods, and people who answered negatively to this question were classified as potential *users* of explanations in safety-critical system application.

Validating Responses. We applied two filtering criteria to eliminate non-reliable or biased answers. While people were allowed to fill in more than one survey, we only took into account the first survey for each person for the main results. This is to remove positively skewed responses resulting from familiarizing with the datasets, further discussed in the results. Additionally, we eliminated participants scoring worse than random, *i.e.* with less than 15 out of 30 correct answers. In total, 2190 answers from 73 participants were analyzed.

Datasets. For the survey, we selected two distinct datasets containing scaled signals of electrical activity and sensor data measured in Volt.

1. **ECG200:** We use the ECG200 (Olszewski (2001)) dataset from the UCR archive (Dau et al. (2019)) containing data of electrical activity measured during one heartbeat. Specifically, the latter part of a heart beat is shown in the signal, starting after the peak point R. The characteristic properties of a normal heart beat (class 1) compared to an ischemic heart beat (class 2) are the high peak point R and the limited recovery time from its minimum S to T. We show a scaled reconstruction of the ground truth in Fig. 4 together with the characteristic points R, S, T, and U.

2. **Artificial Dataset:** Furthermore, we created an artificial dataset, reproducing signals from machine sensors in a noisy environment.In particular, we used four basic time-series shapes, shown in the ground truth signals Fig. 5, and added multiplicative and additive noise with an amplitude of 0 to 1.1, drawn from a uniform distribution.

Black Box Model. We used a Fully Convolutional neural Network (FCN) (Fawaz et al. (2019)) to classify the signals. It consists of three convolutional layers with 128, 256, and 128 filters of kernel size 8, 5, and 3. The first two layers use ReLU activation and batch normalization. The last layer's output is globally averaged and fed into a softmax activation.

3. The study was conducted in compliance with the CERN (2022) Data Privacy Protection Policy and the CERN (2010) Code of Conduct.

Explanation Methods. We compare our MAP explanation method, described in Section 4, with two addition concepts explanation methods. Their implementation details are stated below. The number of concepts is set to two times the number of classes in all datasets $M = 4$, which allows all loss terms to converge. We also tried to keep the number of concepts low and even to improve the simplicity of our survey.

1. **Explanation-By-Example (EBE):** First, we used the idea of instance explanation-by-example methods (Jeyakumar et al. (2020); Papernot and McDaniel (2018)) to implement a global explanation method. Namely, we split the FCN into an encoder $g(\cdot)$ and a predictor, at the last convolutional layer. We then calculate the k-means cluster centers of the activations $g(\mathbf{x}_n)$. The instance with the closest euclidean distance to each cluster center was then used as a global explanation-by-example.
2. **Model-Specific Prototypes (MSP):** We implemented the model-specific prototype method from Gee et al. (2019). This method also learns prototypes from the output of an encoder. A softmax classifier then uses the distance of the encoder output to all prototypes for classification. Finally, the learned prototypes are reconstructed with a decoder. In addition to the cross-entropy loss and the reconstruction loss, the authors introduce a prototype diversity loss as a learning objective. As the method is model-specific, we used the convolutional layers of the FCN as an encoder on top of a fully connected layer with 20 neurons as an encoder $g(\cdot)$. Similar to our MAP model we used a three layer fully connected neural network, with 300 neurons per layer and a sigmoid activation function in the second layer as a decoder $h(\cdot)$. Similarly to the paper (Gee et al. (2019)), the predictor consists of a softmax layer, where decisions are inferred from the distance of input instances to the learned prototypes. All hyperparameters were taken from the original paper, after performing a detailed sensitivity analysis.

Training Stability. While the training of the FCN already converged after 200 epochs, we trained both autoencoder methods for 1500 epochs to ensure convergence of all regularization terms. We ensured that none of the models was stuck in local minimum, by training each model five times and selecting the one with the lowest overall loss.

Study Significance. Confidence intervals are calculated using the binomial proportion (Brown et al. (2001)) $\hat{p} \pm z\sqrt{(\hat{p}(1-\hat{p}))/n}$, where $\hat{p}$ is the proportion of successes in a binomial trial, *i.e.* the amount of all correctly classified instances divided by the amount of all classification samples n. Here, z is the quantile of a standard normal distribution $1 - \alpha/2$, where α is the target error rate. This means that for our 95% confidence interval $\alpha = 0.5$ and $z = 1.96$.

6. Results

6.1. Modeling Results

The quantitative modeling results of EBE, MSP, and MAP are shown in Table 1. Based on the definitions given in Section 3, the classification accuracy, the concept representability, and the reconstructed concept representability are shown by the mean and standard deviation (in brackets) over five training runs.

The EBE & MAP methods, use the same FCN classifier for the prediction of the classes. This FCN classifier is trained only with the cross-entropy loss, without a specific loss for

Table 1: Concept properties of EBE, MSP and our MAP. The model accuracy (higher scores are better), the representabilities (lower scores are better) are given by the mean of five independent training runs with standard deviation in brackets. The accuracy of EBE and MAP is equal, as they use the same FCN classifier for prediction.

Dataset	Model Accuracy [%]		Concept Representability			Reconstructed Concept Representability		
	FCN(ours)	MSP	EBE	MSP	MAP(ours)	EBE	MSP	MAP(ours)
ECG200	**84.0(0.8)**	79.7(10.8)	2.3(0.1)	2.1(1.1)	**0.2(0.1)**	8.1(4.6)	5.8(2.4)	**0.6(0.1)**
Artificial data	**99.9(0.1)**	93.6(17.7)	15.7(0.7)	15.7(4.8)	**7.8(0.8)**	4.7(2.7)	4.7(0.9)	**4.3(1.8)**
ACSF1	**85.9(2.5)**	**85.9(3.1)**	1.0(0.2)	1.4(0.1)	**0.5(0.1)**	1.8(0.9)	2.8(1.9)	**1.2(0.2)**
Computers	**83.5(2.0)**	73.8(11.0)	6.0(0.9)	12.0(4.8)	**1.5(0.2)**	**2.1(0.0)**	13.2(0.5)	9.3(1.0)
ECG5000	**92.9(0.2)**	92.7(0.7)	3.9(0.2)	13.0(1.5)	**2.0(0.3)**	8.9(0.3)	**5.6(1.3)**	6.3(0.5)
LargeKitchenAppliances	86.9(0.6)	**89.4(1.2)**	9.5(0.5)	15.5(0.3)	**1.3(0.5)**	15.5(1.2)	11.8(2.5)	**9.1(0.4)**
PowerCons	**91.9(0.8)**	78.7(19.1)	2.7(0.1)	4.7(2.2)	**1.2(0.2)**	9.6(1.5)	**5.3(2.6)**	7.4(1.7)
RefrigerationDevices	50.1(1.8)	**50.3(3.1)**	5.4(0.3)	11.5(0.7)	**4.1(1.0)**	**12.6(0.6)**	14.2(1.0)	18.3(1.4)
ScreenType	61.8(1.8)	**62.1(2.4)**	5.0(0.2)	10.9(0.9)	**1.8(0.2)**	**7.2(0.8)**	14.7(2.9)	12.4(1.1)
SmallKitchenAppliances	**78.5(1.2)**	74.7(3.3)	8.8(0.3)	14.9(1.0)	**1.5(0.7)**	**13.3(0.1)**	18.6(0.8)	15.5(0.6)
Plane	**99.4(1.1)**	95.6(7.7)	0.4(0.1)	0.9(0.8)	**0.2(0.1)**	1.0(0.3)	2.8(1.9)	**0.9(0.5)**
Trace	**100.0(0.0)**	**100.0(0.0)**	1.2(0.0)	1.9(0.1)	**0.1(0.0)**	2.6(0.9)	2.0(0.2)	**0.9(0.2)**
Win	9	5	0	0	12	4	2	6

explanation. The MSP method is trained to classify and explain at the same time. As a result of this combined objective, the model does not always converge to the global minimum of the cost function. This effect is also observed in the Artificial data, the Computers, the PowerCons, and the Plane dataset, where the standard deviation of the MSP is much higher compared to the standard deviation of the FCN. If the MSP does converge, then it reaches similar results compared to the FCN. In case of the LargeKitchenAppliances, the ScreenType, and the SmallKitchenAppliances datasets, the mean accuracies of MSP are even higher compared to the FCN mean accuracy.

The MAP reaches the highest concept representability in all cases. This shows that the distribution of the latent space is representing the input distribution most accurately for the MAP. While the MAP derives the latent space with a linear transformation, the decoder is still able to identify the correct concepts. This can be seen in the high reconstructed concept representability, where the MAP achieves best results for six datasets. The reconstructed concept representability of EBE is highest in four datasets. The similarity between input instances and explanation-by-example concepts is more similar for EBE compared to prototypes of MSP and MAP, where unimportant information, *e.g.* noise, is filtered out.

We further obtained the true reconstructed concept representability on the artificial dataset, where the ground truth is available. Here, we used the ground truth signal of each input instance as a concept. For this case we obtain a reconstructed concept representability of 3.7, while MSP and EBE reaches 4.7, and MAP reaches 4.3. Looking at Figure 5, the prototypes of MAP are closest to the ground truth, which validates the performance measure.

6.2. Human User Study Results

The results of the study including 73 Participants which classified a total of 2190 instances are presented in Table 2. Analyzing the answers of all participants from both datasets, our MAP method, showed the best results with 79.3% correct answers. This observation is valid also when taking into account the non-overlapping confidence intervals. When looking at the same quantity for individual datasets, one can observe a similar trend.

Figure 4: Ground truth and explanation of the ECG200 (Olszewski (2001)) dataset, showing the latter part of a heart beat, starting before the peak R. For each class, two concepts were extracted with different explanation methods.

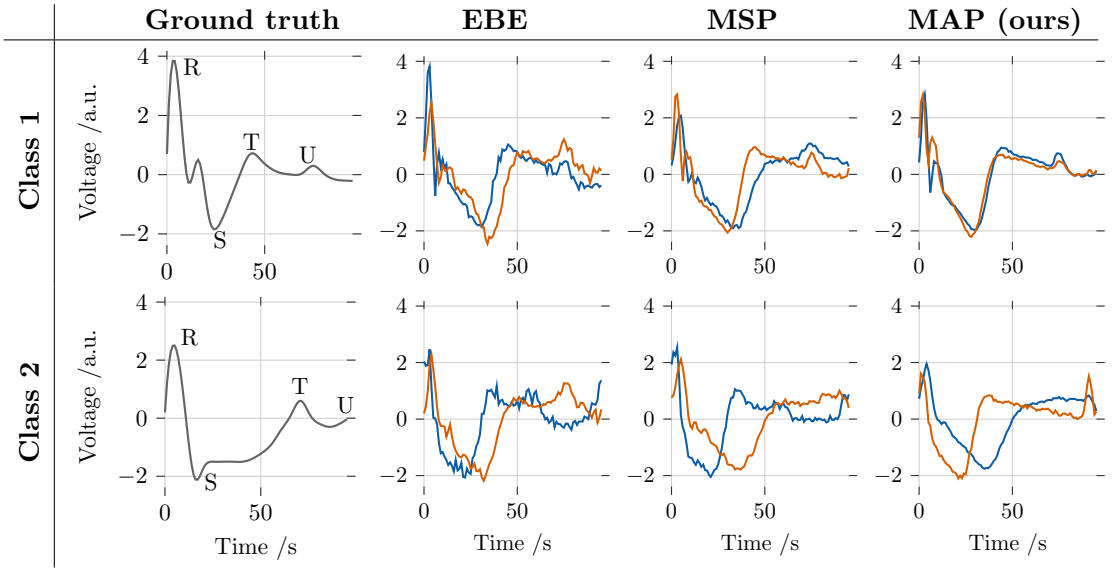
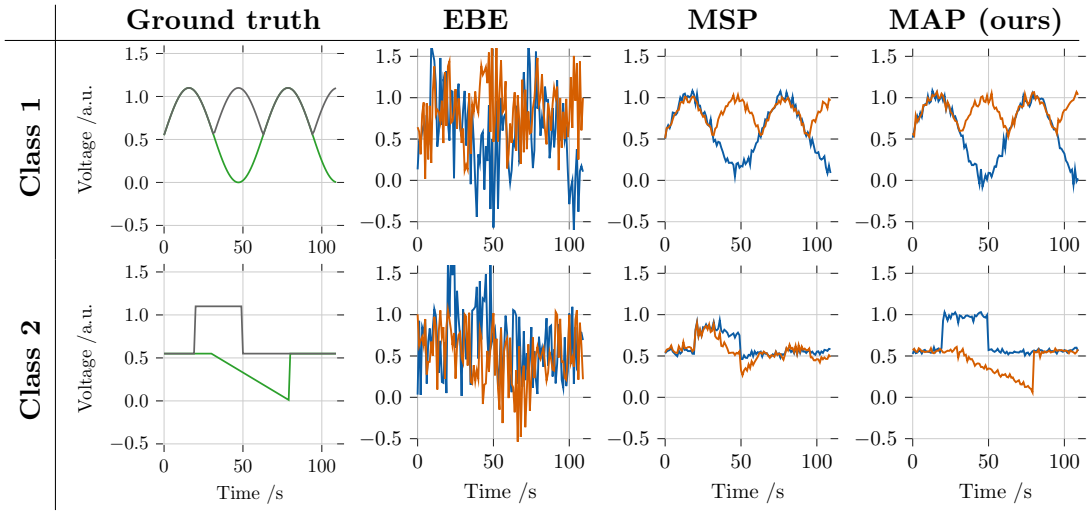


Figure 5: Explanations of artificially created dataset with two concepts per class, extracted from different explanation methods. The ground truth signal shows the four shapes within the dataset, to which multiplicative and additive noise with an amplitude of 0 to 1.1, drawn from a uniform distribution, was added.



In the artificial dataset, EBE was preferred by *users* over other methods. When considering Fig. 5, one would not expect this, as it seems that EBE is the most distinct from the ground truth signal. However, possibly participants not familiar with machine learning, were not able to establish the link between the pattern in the abstract concepts and the

Table 2: Results of the study comparing global EBE, MSP and our MAP. 73 Participants classified a total of 2190 instances, given the reconstructed concepts $\hat{\mathbf{x}}$ and their labels $\hat{y}$. The classification accuracy denotes their receptivity r (see Eq. 1) for the calculated concepts and is shown with the 95% confidence interval for a binomial proportion.

Participants	**Method**	**ECG200** [%]	**Artificial data** [%]	**Total** [%]
Developer	EBE	65.6 ± 6.7	79.0 ± 5.7	72.3 ± 4.4
	MSP	64.4 ± 8.1	74.1 ± 7.4	69.3 ± 5.5
	MAP (ours)	**74.8 ± 5.2**	**84.8 ± 4.3**	**79.8 ± 3.4**
User	EBE	71.3 ± 7.3	**80.0 ± 6.4**	75.7 ± 4.9
	MSP	64.3 ± 6.5	67.6 ± 6.3	66.0 ± 4.5
	MAP (ours)	**77.8 ± 7.0**	78.5 ± 6.9	**78.1 ± 5.0**
All participants	EBE	68.1 ± 4.1	79.4 ± 3.1	73.8 ± 2.4
	MSP	64.3 ± 4.0	70.1 ± 3.2	67.2 ± 2.4
	MAP (ours)	**75.8 ± 3.6**	**82.7 ± 2.9**	**79.3 ± 2.2**

time-series instances. Looking at the performance of the MAP method on the artificial dataset, prototypes that did not fit the shape of the ground truth, appear to be confusing for the *users*.

For the ECG200 dataset, *developers* and *users* were able to generalize best using our method with 74.8% and 77.8% correct answers, respectively. Looking at the class 2 signal in Fig. 4, the characteristic features of ischemic heartbeat signals are represented well by the derived concept. Specifically, the low amplitude in the spike R and the long recovery time from the points S to T is visualized, while showing much less noise than the other methods. A trend of *developers* giving worse results than *users* is visible, suggesting that *developers* are not necessarily able to generalize better than *users* utilizing concept-based explanations.

We further evaluated the effect of our filtering criteria (see Section 5.2), by looking at the results of the 1170 dropped answers from 39 participants who filled out more than one survey. Here, the learning effect outweighed the decision fatigue, as the performance increased on average by 6.8% for ECG200 and 5.1% for the artificial dataset in later attempts.

7. Conclusion

The quality of global, and model-agnostic concept explanation techniques is a key factor to help experts in safety critical domains gaining trust in predictions made by machine learning models. We demonstrated that our provided model-agnostic method fulfills these requirements by providing accurate and complete explanations, independent of the weight initialization or the concept numbers. We assessed the quality of our explanations quantitatively with 12 datasets, containing data common in safety critical applications. On two datasets, we further performed a human user study across 75 participants with, 2190 validated answers. The conducted survey showed that our proposed method helped participants to generalize explanations for classification tasks on time-series data across all datasets and target audiences. Specifically, participants reached 79.3% correct answers on average using our method, while reaching only 73.8% with explanation-by-example and 67.2% with model-specific prototypes. In the case of the artificial dataset, the prototype explanations show a significant visual discrepancy with respect to the signals presented in the survey, possibly leading to better results of the explanation-by-example method. In our domain,

i.e. predicting failures in particle accelerators, explanations are expected to be interpreted by domain experts. This makes the explanation by model-agnostic prototypes the preferred option in general, with explanation-by-example representing a valid alternative if prototypes become too abstract.

8. Future Work

Our future work will focus on the application of the proposed method to predict failures in superconducting electrical circuits in CERN's Large Hadron Collider (LHC Wenninger (2016)). The circuit data collected during several years of successful operation enables the use of data-driven methods to help experts find anomalies in the behavior of superconducting circuits and potentially also of protection systems. Our model-agnostic explanation technique will help in explaining existing deep learning models to system experts with no machine learning background. This will enable faster and more accurate fault diagnostics and optimized maintenance actions, further increasing safety and availability of the LHC. In addition, improvements of our method will aim at making more tailored variants of our explanation. Particularly, we plan to use Fourier analysis to correctly address the complexity of the behavior of superconducting circuits in the frequency domain.

References

Abien Fred Agarap. Deep learning using rectified linear units. *CoRR*, abs/1803.08375, 2018.

Ghassan AlRegib and Mohit Prabhushankar. Explanatory paradigms in neural networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, 39(4): 59–72, 2022.

Sebastian Bach et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one*, 10(7):1–46, 2015.

Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.

José Manuel Beítez, Juan Luis Castro, and Ignacio Requena. Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, 8(5):1156–1164, 1997.

Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, 2011.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101–133, 2001.

CERN. Code of conduct, 2010. URL https://procurement.web.cern.ch/system/files/document/cern-code-conduct.pdf.

CERN. Data privacy protection policy, 2022. URL https://home.cern/data-privacy-protection-policy.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter conference on applications of computer vision*. IEEE, 2018.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*. NeurIPS, 2019.

Subhajit Das, Panpan Xu, Zeng Dai, Alex Endert, and Liu Ren. Interpreting deep neural networks through prototype factorization. In *International Conference on Data Mining Workshops*. IEEE, 2020.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Hassan Ismail Fawaz et al. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

Alan H Gee, Diego Garcia-Olano, Joydeep Ghosh, and David Paydarfar. Explaining Deep Classification of Time-Series Data with Learned Prototypes. *CoRR*, abs/1904.0, 2019.

Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. Explaining any time series classifier. In *Second International Conference on Cognitive Machine Intelligence*. IEEE, 2020.

Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz*, 2020.

Jeya V. Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. In *Advances in Neural Information Processing Systems*. NeurIPS, 2020.

Been Kim et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *International conference on ML*. PMLR, 2018.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep Learning for Case-based Reasoning through Prototypes. *CoRR*, abs/1710.04806, 2017.

Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data*, 12(5):1–35, 2018.

Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. NeurIPS, 2017.

Rami Mochaourab, Arun Venkitaraman, Isak Samsten, Panagiotis Papapetrou, and Cristian R Rojas. Post-hoc explainability for time series classification: Toward a signal processing perspective. *IEEE signal processing magazine*, 2022.

Christoph Obermair et al. Explainable Machine Learning for Breakdown Prediction in High Gradient RF Cavities. *CoRR*, abs/2202.05610, 2022.

Robert Thomas Olszewski. *Generalized feature extraction for structural pattern recognition in time series data*. PhD thesis, Ann Arbor, 2001.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.

Nicolas Papernot and Patrick D McDaniel. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *CoRR*, abs/1803.0, 2018.

Sara Reardon. Rise of robot radiologists. *Nature*, 576(7787):54–54, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR*, abs/1602.04938, 2016.

Thomas Rojat et al. Explainable artificial intelligence (XAI) on timeseries data: A survey. *CoRR*, abs/2104.0, 2021.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Interpretable Deep Learning by Propagating Activation Differences. *CoRR*, abs/1605.01713, 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

Wensi Tang, Lu Liu, and Guodong Long. Interpretable time-series classification on few-shot samples. In *International Joint Conference on Neural Networks*. IEEE, 2020.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence: Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(86):2579–2605, 2008.

Stephen F Weng et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):1–14, 2017.

Jörg Wenninger. Machine Protection and Operation for LHC. *Proceedings of the Joint International Accelerator School*, 2(1):377–401, 2016.

Chih-Kuan Yeh et al. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. In *Advances in Neural Information Processing Systems*. NeurIPS, 2020.

Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. Tapnet: Multivariate time series classification with attentional prototypical network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6845–6852, 2020.

Paper 6

Interpretable Anomaly Detection in the LHC Main Dipole Circuits with Non-negative Matrix Factorization

Interpretable Anomaly Detection in the LHC Main Dipole Circuits with Non-Negative Matrix Factorization

Christoph Obermair, Andrea Apollonio, Zinour Charifoulline, Lukas Felsberger, Marvin Janitschke, Franz Pernkopf, Emmanuele Ravaioli, Arjan Verweij, Daniel Wollmann, and Mariusz Wozniak

***Abstract*—CERN's Large Hadron Collider (LHC) with its eight superconducting main dipole circuits has been in operation for over a decade. During this time, relevant operational parameters of the circuits, including circuit current, voltages across magnets and their coils, and current to ground, have been recorded. These data allow for a comprehensive analysis of the circuit characteristics, the interaction between their components, and their variation over time. Such insights are essential to understand the state of health of the circuits and to detect and react to hardware fatigue and degradation at an early stage.**

In this work, a systematic approach is presented to better understand the behavior of the main LHC dipole circuits following fast power aborts. Non-negative Matrix Factorization is used to model the recorded frequency spectra as common sub-spectra, by decomposing the recorded data as a linear combination of basis vectors, which are then related to hardware properties. The loss in reconstructing the recorded frequency spectra allows to distinguish between normal and abnormal magnet behavior. In the case of abnormal behavior, the analysis of the sub-spectra properties enables to infer possible hardware issues. Following this approach, five dipole magnets with abnormal behavior were identified, of which one was confirmed to be damaged. As three of the other four identified magnets share similar sub-spectra characteristics, they are also treated as potentially critical. These results are essential for preparing targeted magnet measurements and may lead to preventive replacements.

***Index Terms*—Large Hadron Collider, Quench Protection, Non-negative Matrix Factorization, Machine Learning**

I. Introduction

THE LHC is the world's highest energy particle accelerator, relying on 1232 superconducting main dipole magnets to bend the high-energy particle beams along its circumference. These dipole magnets are powered through eight separate circuits of 154 magnets each. To reach the nominal field of 8.0 T and a current of 11.85 kA, each magnet is cooled down to 1.9 K with superfluid helium. At this temperature, the magnet is superconducting. A resistive transition in a superconducting magnet, also called quench, results in local heating in the superconducting cables and high voltage transients in the magnet, which can possibly damage the magnet if not appropriately managed. In case of a quench or other powering failures in the circuits, a system of protection elements is in place to safely dissipate the energy in the quenched magnets and extract the remaining energy from the circuit [1]. This process is referred to as a Fast Power Abort (FPA) event. To better understand the data recorded during a FPA event, the LHC main dipole circuits and their protection system are explained in more detail below.

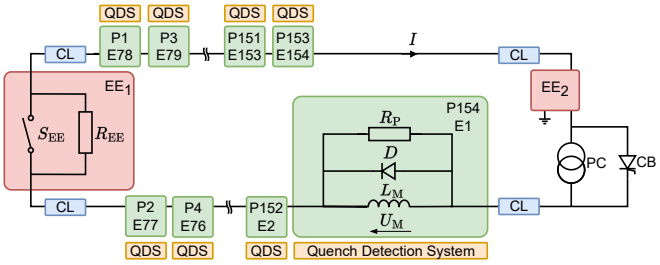


Fig. 1. Schematic view of the main dipole circuit, including the Power Converter (PC), the Crowbar (CB) and the Current Leads (CL). The Quench Detection System (QDS) triggers a Fast Power Abort (FPA), which deactivates the PC and activates the energy extraction systems. Furthermore, it triggers the discharge of the quench heaters (QH) in the respective magnet, if a quench is detected. The two Energy Extraction systems EE_1 and EE_2 consist of a switch S_{EE}, and an Energy Extraction Resistance R_{EE}. The circuit is grounded at the center of the resistor R_{EE} in the EE_2 system. The magnet with inductance L_M and the by-pass Diode D with a Parallel Resistance R_P are in a liquid helium cryostat. Magnets are labeled by their Physical position (P) from the left to the right. The Electrical positions (E) are counted clockwise along the electrical connection starting from the PC. The numbering shown here is representing the circuits in sectors 12, 34, 56, and 78. In sectors 23, 45, 67, and 81 the electrical labels are inverted, as the PC is on the left side of the circuit.

Figure 1 shows a schematic view of a main dipole circuit with its 154 magnets, each represented by a magnet inductance L_M [2]. For this analysis, the magnets are counted along their physical position from the left to the right or clockwise along the electrical connection starting from the Power Converter (PC). In case the PC is switched off, the current I circulating in the circuit by-passes the PC via the Crowbar (CB). The Current Leads (CL) indicate the transition between the cold superconducting part of the circuit and the warm, normal conducting part of the circuit.

The protection system includes a Quench Detection

Christoph Obermair is with CERN, CH, 1211 Meyrin, Switzerland, and also with Graz University of Technology, AT, 8010 Graz, Austria (e-mail: christoph.obermair@cern.ch).

Andrea Apollonio, Zinour Charifoulline, Lukas Felsberger, Emmanuele Ravaioli, Arjan Verweij, Daniel Wollmann, Mariusz Wozniak are with CERN, CH, 1211 Meyrin, Switzerland.

Marvin Janitschke is with CERN, CH, 1211 Meyrin, Switzerland, and also with the University of Rostock, DE, 18051 Rostock, Germany.

Franz Pernkopf is with Graz University of Technology, AT, 8010 Graz, Austria.

The code for this project is publicly available at https://github.com/cobermai/lhc_anomaly_detection

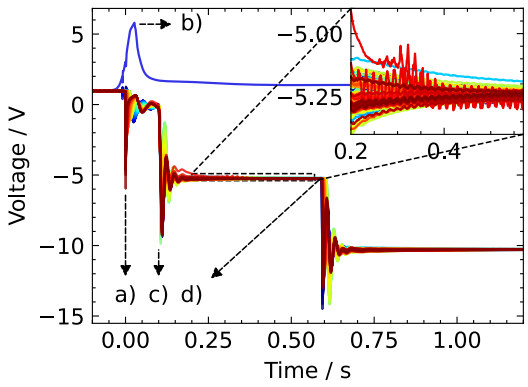


Fig. 2. Voltages U_{M} across the 154 main dipole magnets of sector 78 following a quench in the magnet with the electrical position 141 on 31.03.2021 with its different phases: a) the FPA is triggered at 0 s, the QHs for magnet 141 are activated, and the PC is deactivated; b) after around 0.03 s the by-pass diode of the quenched magnet becomes conductive; c) the first Energy Extraction system EE_1 is activated about 0.1 s after the FPA trigger; d) the second Energy Extraction system EE_2 is activated approximately 0.5 s after the first one. The blue curve shows the voltage across the quenched magnet, while the remaining curves represent the voltages across the other 153 magnets of the circuit. The voltage signals of one of the analyzed plateaus are shown in the magnified view on the right top corner.

System (QDS), which detects a quench and triggers the appropriate protection actions [3], [4]. Upon the detection of a quench in a dipole magnet, the PC is switched off and the Quench Heaters (QH) of this magnet are activated. QHs are resistive strips attached to the outer surface of each magnet coil [5]. They ensure protection by distributing the magnet's stored energy more uniformly over the quenched magnet windings [6]. The by-pass Diode D diverts current from the quenched magnet. This restricts the quenching magnet to only absorb its stored magnetic energy, not the energy of the entire circuit. The Parallel Resistance R_{P} installed across each magnet, smoothens transient voltages during this process [7]. To avoid the circuit's energy to solely discharge in the diode of the quenched magnet, the switches S_{EE} in both Energy Extraction (EE) systems are sequentially activated [8]. They direct the circuit current towards the Resistances R_{EE}, which extracts the circuit's energy within around 300 s. The voltages U_{M} measured over the 154 magnets during a FPA event are shown in Fig. 2. These signals are voltage transients that contain information about the behavior of the electrical circuit and its components [9].

A quench is a routine occurrence, during training periods aimed at increasing the peak magnetic field in the superconducting magnets [10], and rarely occurs during operation. Once a quench emerges, it is frequently accompanied by secondary quenches. Secondary quenches result from electromagnetic perturbations milliseconds after the initial quench [2] or from thermal propagation in the helium tens of seconds later [11]. Following a quench or a secondary quench, the magnet is exposed to local heating, high voltages, and thermal expansion, depending on the circuit's energy level [12]. The magnets are designed with additional margins for this case, but defects can still occur. Certain hardware failures in the superconducting circuits can notably impact the availability of the LHC, potentially resulting in months of downtime. The understanding of normal and abnormal circuit dynamics helps to ensure safe quench mitigation and to detect precursors of hardware failures, allowing to schedule preventive maintenance.

To better understand the circuit dynamics, local frequency responses of selected main dipole magnets have been measured and evaluated [13], [14]. FPA events have been deliberately triggered in all main dipole circuits to better understand the voltage transients in the circuits in the absence of a quench [15]. The main dipole circuits have also been extensively studied with electrical simulations with Simulation of Transient Effects in Accelerator Magnets (STEAM) framework [16], [17].

While these simulations account for much of the circuit's behavior and measurements, some aspects - such as the voltage transients observed in the magnified view of Fig. 2 following activation of the EE systems - cannot be captured entirely by simulation. In this time window, secondary quenches frequently occur due to electromagnetic perturbations [2]. A better understanding of the circuit behavior, and in particular the voltage transients after triggering the EE systems, allows the development of mitigation strategies to reduce the number of these electromagnetically induced quenches and the risk associated with them.

The presented research aims to provide insights into the propagation and physical process explaining the observed frequency spectra of the magnet voltage after activation of the EE systems. Normal and abnormal behavior in these frequency spectra is detected and characterized.

The detection of normal and abnormal behavior and their associated physical processes is carried out by Non-negative Matrix Factorization (NMF). The choice of NMF is motivated by recent successful applications of data-driven models to predict quenches [18], to classify QH failures [19], or to model the voltage across magnets [20].

NMF aims at providing interpretable results, as the lack of interpretability is a frequent criticism to other data-driven methods [21], [22]. The method was originally used to decompose pictures of human faces into coherent components like eyes, mouth etc. [23]. The decomposed components are additive and are therefore easy to understand by humans. NMF has been successfully applied to discover molecular patterns in genes [24], to separate different sources of a mixed acoustic signal [25], and to derive properties of galaxies from astronomical observations [26]. In the context of this research, NMF is used to decompose the frequency spectra of the voltages recorded in the LHC's main dipole circuits during FPA events (see Fig. 2) and understand the physical processes causing them. The loss introduced by the frequency decomposition for each FPA event allows detecting and interpreting abnormal behavior in the circuits.

The remainder of the paper is structured as follows. In Section II, an overview of NMF within the context of this study is given. In Section III, the results are presented by showing possible causal relationships between the distinct frequencies in the circuit and the circuit hardware. In addition, five abnormal FPA events are highlighted, and their

characteristic frequencies are interpreted. In Section III-E, the risk of abnormal FPA events for the different hardware components of the machine is elaborated. Finally, Section IV summarizes the results.

II. Methodology

A. Available Data and Preprocessing

This subsection explains the selection and pre-processing of the measured voltage signals. After the FPA is triggered (point a. in Fig. 2), the first EE system is activated 0.1 s later (Fig. 2 b.). The second EE system is activated after further 0.5 s (Fig. 2 d.). The two periods analyzed in this study are the two voltage plateaus [0.2; 0.575] and [0.7; 1.075] seconds after the triggering of the FPA. These were chosen because the observed frequency spectra in the magnet voltages are not reconstructed by the existing simulation models, which are based on the current knowledge of the circuit's behaviour. Each of these plateaus covers $P = 154$ voltage signals recorded with a sampling rate of 1068 Hz over a length of 0.375 s.

All data used for this study have been recorded after 2017, as the activation times of the EE systems have been kept unchanged since then. In total $Q = 699$ distinct FPA events have been used. These events are split into three categories: 48 events do not contain a quench, 494 events contain a single quench of one magnet in the circuit, and 157 events contain at least one secondary quench due to electromagnetic perturbations. The frequency spectra of the latter events deviate strongly from the others, therefore, only the 48 events without a quench and the 494 events with a single quench are compared to derive anomalies in Subsection II-E. In contrary, the spectral components of all $Q = 699$ events are interpreted in Subsection II-D. Secondary quenches due to thermal propagation in the helium do not affect the frequency spectra in the voltage plateaus, as they occur at a later stage. Events with those secondary quenches are treated like events with a single quench.

The 699 events with voltage signals from 154 magnets for each of the two plateaus after the activation of the EE systems yield a total of $M = 2 \cdot P \cdot Q = 215292$ distinct voltage signals. Each of these $M = 215292$ voltage signals is transformed into a frequency spectrum with $N = 200$ data points via a Fast Fourier Transformation (FFT), an efficient algorithm for computing the discrete Fourier transform [27]. The Nyquist criterion allows showing frequencies of up to 534 Hz [28]. In order to mitigate spectral leakage seven window functions including a window-specific amplitude correction are compared. The window functions are: Rectangular, Hanning, Hamming, Bartlett, Blackman, Flat-top, and Tukey [29]. An exponential trend $\bar{\mathbf{x}}$ of the form

$$\bar{\mathbf{x}} = Ae^{-\mathbf{t}/\tau} + C, \tag{1}$$

is fitted with least squares [30] and subtracted from each individual voltage signal $\mathbf{x}$ with timestamps $\mathbf{t}$. A corresponds to the amplitude of the decay, τ to the decay's time constant, and C to the offset. This exponential trend, which is best visible in the voltage signal with the highest amplitude in the magnified view of Fig. 2, corresponds to non-linear effects in the magnets [2]. Applying these pre-processing steps yields a dataset composed of M frequency spectra from Q FPA events, with N data points in each frequency spectrum that is processed by NMF.

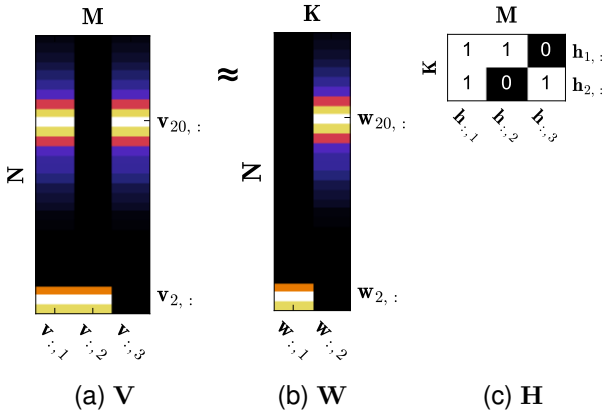


Fig. 3. Example of the NMF decomposition $\mathbf{V} \approx \mathbf{WH}$. In this example, $M = 3$ frequency spectra with $N = 30$ data points are approximated with $K = 2$ spectral components. The color spectrum indicates the voltage amplitudes and ranges from black 0 to white 1. The vertical index count i starts at the bottom.

B. Non-negative Matrix Factorization

Using N and M defined above, let $\mathbf{V}$ be the input matrix, with entries $v_{i,j}$ for $i = 1,..,N$ and $j = 1,..,M$. NMF decomposes the $N \times M$ matrix $\mathbf{V}$ into a product of a $N \times K$ matrix $\mathbf{W}$ and a $K \times M$ matrix $\mathbf{H}$ such that:

$$\mathbf{V} \approx \mathbf{WH}. \tag{2}$$

Here, $\mathbf{W}$ represents the spectral components and $\mathbf{H}$ their weights. The parameter K defines the number of spectral components. All elements $w_{i,k}$ and $h_{k,j}$ of the matrices $\mathbf{W}$ and $\mathbf{H}$ are constrained to be non-negative, leading to the additive nature of the NMF decomposition [23].

Figure 3 illustrates this behavior for a simplified example: the $K = 2$ spectral components can be added to reconstruct the $M = 3$ frequency spectra. For simplicity, the shorthand representation ":" is used to identify all entries from one dimension, *e.g.* $[\mathbf{W}]_{i=[1,..,N],k=1} = \mathbf{w}_{:,1}$. The two spectral components $\mathbf{w}_{:,1}$ and $\mathbf{w}_{:,2}$, have their maximum of 1 at $i = 2$ and $i = 20$, respectively.

Optimizing $\mathbf{W}$ and $\mathbf{H}$ involves minimizing an element-wise similarity metric $d_*(\cdot)$ between the input $v_{i,j}$ and the reconstructed input $\hat{v}_{i,j} = \sum_k^K w_{i,k}h_{k,j}$. Three widely used similarity metrics are:

1) Squared Eucleadian (Eu) distance [31]:

$$d_{Eu}(v_{i,j}, \hat{v}_{i,j}) = \|v_{i,j} - \hat{v}_{i,j}\|^2 \tag{3}$$

2) Generalized Kullback-Leibler (KL) divergence [32]:

$$d_{KL}(v_{i,j}, \hat{v}_{i,j}) = v_{i,j}\log\frac{v_{i,j}}{\hat{v}_{i,j}} - v_{i,j} + \hat{v}_{i,j} \tag{4}$$

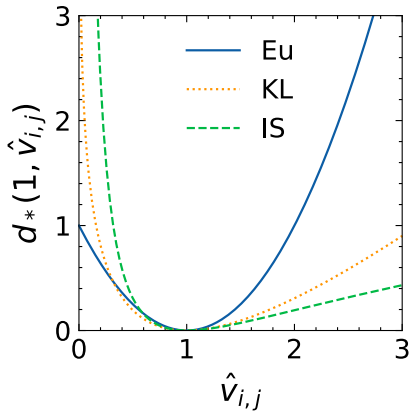


Fig. 4. The Squared Eucleadian (Eu) distance, the Generalized Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence as a function of $\hat{v}_{i,j}$, assuming $v_{i,j} = 1$.

3) Itakura-Saito (IS) divergence [33]:

$$d_{IS}(v_{i,j}, \hat{v}_{i,j}) = \frac{v_{i,j}}{\hat{v}_{i,j}} - \log\frac{v_{i,j}}{\hat{v}_{i,j}} - 1 \tag{5}$$

Figure 4 shows the three metrics as a function of $\hat{v}_{i,j}$ with $v_{i,j} = 1$. The Eu-distance squares the absolute difference, demonstrated by this example: $d_{Eu}(1,2) = d_{Eu}(100,101)$. Figure 4 (Eu) therefore shows a quadratic function, equally sensitive to $\hat{v}_{i,j}$ greater or less than one. The KL-divergence reflects the relative entropy, corresponding to the energy in a system. This causes an increased sensitivity for under-estimation and a decreased sensitivity for over-estimation of the reconstruction $\hat{v}_{i,j}$ [32]. This is reflected in Fig. 4 (KL) by a larger distance metric for $\hat{v}_{i,j} < 1$ as compared to $\hat{v}_{i,j} > 1$. The IS-divergence is scale-invariant [34] as it compares relative differences, illustrated by $d_{IS}(1,2) = d_{IS}(100,200)$. The effect of sensitivity observed for the KL-divergence is amplified for the IS-divergence, as is visible in Fig. 4 (IS) [33]. The advantages and weaknesses of these properties will be discussed in Section III-A. Using $d_*(\cdot)$, the reconstruction loss is obtained by $\sum_i^N \sum_j^M d_*(v_{i,j}, \hat{v}_{i,j})$.

All values of $\mathbf{W}$ and $\mathbf{H}$ are initialized with average non-negative double Singular Value Decomposition (SVD) [35]. The term "double" is derived from the use of SVD in approximating both matrices $\mathbf{W}$ and $\mathbf{H}$. This method leads to faster convergence and more robust spectral components compared to random initialization. Any zero values derived by SVD are replaced by the global average of $\mathbf{V}$ as they would otherwise remain at zero during the consecutive multiplicative updates. These multiplicative updates of $w_{i,k}$ and $h_{k,j}$ are specific to each of the three similarity measures [25]:

4) Squared Eucleadian (Eu) distance:

$$w_{i,k} \leftarrow w_{i,k}\frac{\sum_j v_{i,j}h_{k,j}}{\sum_j \hat{v}_{i,j}h_{k,j}},\ h_{k,j} \leftarrow h_{k,j}\frac{\sum_i v_{i,j}w_{i,k}}{\sum_i \hat{v}_{i,j}w_{i,k}} \tag{6}$$

5) Generalized Kullback-Leibler (KL) divergence:

$$w_{i,k} \leftarrow w_{i,k}\frac{\sum_j \frac{v_{i,j}}{\hat{v}_{i,j}}h_{k,j}}{\sum_j h_{k,j}},\ h_{k,j} \leftarrow h_{k,j}\frac{\sum_i \frac{v_{i,j}}{\hat{v}_{i,j}}w_{i,k}}{\sum_i w_{i,k}} \tag{7}$$

6) Itakura-Saito (IS-divergence) divergence:

$$w_{i,k} \leftarrow w_{i,k}\sqrt{\frac{\sum_j \frac{v_{i,j}}{\hat{v}_{i,j}}\frac{h_{i,j}}{\hat{v}_{i,j}}}{\sum_j \frac{h_{i,j}}{\hat{v}_{i,j}}}},\ h_{i,k} \leftarrow h_{i,k}\sqrt{\frac{\sum_i \frac{v_{i,j}}{\hat{v}_{i,j}}\frac{w_{i,j}}{\hat{v}_{i,j}}}{\sum_i \frac{w_{i,j}}{\hat{v}_{i,j}}}}, \tag{8}$$

No NMF regularization [36], [37] is applied to avoid the risk of regularizing spectral components with small amplitudes and to minimize the number of parameters to be optimized that would unavoidably be added with regularization.

C. Spectral Component Identification

In this subsection the methodology to derive the final spectral components $\mathbf{W}$, their number K, and their corresponding weights $\mathbf{H}$ is described. 19 different numbers of spectral components are investigated ($K = 2, ..., 20$). The exact spectral components to which the NMF algorithm converges, also depend on the choice of the three distance measures from Eq. 3-5, and the seven types of distinct window functions. All three parameters are referred to as hyperparameters in the remainder of this paper. In total, 399 possible combinations of parameters exist.

The number of spectral components K determines the resolution of the factorization. Choosing a larger K results in a reduced reconstruction loss. However, in the context of this project, separating the measured frequency spectra into common spectral components aims at representing different physical processes. Hence, more spectral components are only desirable if they can be mapped to separate physical processes. Ideally, one physical process should be represented by one spectral component. To choose K accordingly, an additional performance measure is introduced, based on prior research [39], [40].

This performance measure calculates the mean pairwise Chebyshev distance between column pairs of spectral components and column pairs of their weights. The Chebyshev distance shows the maximum value of the absolute differences between two vectors [38]. For the spectral components, the average Chebyshev distance over all $(K-1)!$ possible pairs of spectral components is used as the performance measure $\bar{d}_{Ch}$. An example to calculate $\bar{d}_{Ch}$ for the spectral components in Fig. 3b is shown below.

	$\mathbf{w}_{:,1}$	$\mathbf{w}_{:,2}$	$\mathbf{w}_{:,1}$	$\mathbf{w}_{:,2}$	$\mathbf{w}^{\text{new}}_{:,1}$
$i = 20$	1	0	0	1	0
$i = 2$	0	1	1	0	1
	$\bar{d}_{Ch} = 1$		$\bar{d}^{\text{new}}_{Ch} = \frac{1+1+0}{3} = \frac{2}{3}$		

Since $\mathbf{W}$ has only two columns, there is one possible column pair of spectral components, resulting in $\bar{d}_{Ch} = \max(|\mathbf{w}_{:,1} - \mathbf{w}_{:,2}|) = 1$. If the columns in Fig. 3b are subtracted, this is evident as their absolute difference is $w_{2,1} = w_{20,2} = 1$. Suppose K is increased by one, and the additional component $\mathbf{w}^{\text{new}}_{:,1}$, happens to be identical to $\mathbf{w}_{:,1}$. In this case $\mathbf{w}^{\text{new}}_{:,1}$ represents the same physical process as $\mathbf{w}_{:,1}$. Consequently, $\max(|\mathbf{w}_{:,1} - \mathbf{w}^{\text{new}}_{:,1}|)$ is zero, leading to a decreased $\bar{d}^{\text{new}}_{Ch}$ on the right side of the example calculation above. This example shows that adding more spectral components K, which are not expected to come from different physical processes, gets penalized by the introduced performance metric.

The performance measure is derived similarly for the spectral component weights $\mathbf{H}$ in Fig. 3c:

	$\mathbf{h}_{:,1}$	$\mathbf{h}_{:,3}$	$\mathbf{h}_{:,3}$	$\mathbf{h}_{:,1}$	$\mathbf{h}_{:,3}$	$\mathbf{h}_{:,3}$
$k=1$	1	1	0	0.5	0.5	0
$k=2$	1	0	1	1	0	1
$k^{\text{new}}=1$	–	–	–	0.5	0.5	0
	$\bar{d}_{Ch}=\frac{1+1+1}{3}=1$			$\bar{d}_{Ch}^{\text{new}}=\frac{1+1+0.5}{3}=\frac{5}{6}$		

For the existing $\mathbf{H}$ in Fig. 3c this is shown on the left side of the example calculation. There are three possible column pairs where the Chebyshev distance of *e.g.* the first column pair is calculated by $\max(|\mathbf{h}_{:,1}-\mathbf{h}_{:,3}|)=1$. If the spectral component $\mathbf{w}_{:,1}^{\text{new}}$ is added, the spectral component weights are adjusted to obtain the same reconstructions. This is illustrated by the right side of the example calculation above. The Chebyshev distance of $\max(|\mathbf{h}_{:,1}-\mathbf{h}_{:,3}|)$ is reduced to 0.5, affecting the average $\bar{d}_{Ch}^{\text{new}}=\frac{5}{6}$ accordingly. Again, the performance measure indicates that K should not be increased. For computational efficiency, this performance metric is not evaluated for all $(M-1)!$ combinations of $M=215292$ frequency spectra, but for $M^*=1000$ randomly chosen frequency spectra. In addition to the performance metric, the final choice of spectral components is based on manual inspection of the identified components. This will be discussed in Subsection III-A.

D. Spectral Component Interpretation

This subsection describes the method of identifying the physical process behind a spectral component. For this purpose, the location of the maximum, the average of the maximum amplitude, and the weight propagation of the spectral components are analyzed and discussed. These allow relating a spectral component to hardware behavior in the LHC main dipole circuits [41].

For each FPA event q, the location of the maximum is defined as the magnet position with the highest of the $P=154$ weights of a spectral component k. The average weight at this position over a selection of FPA events is defined as the average maximum amplitude. A distinction is made between maxima near the quenched magnet, the PC, or the two EE systems.

From the magnet position at which the maximum occurs, the spectral component can propagate to its physical or electrical neighbors. This results from the fact that physical magnet neighbors can experience electromagnetic coupling due to gaseous helium flow between adjacent cryostats, or instrumentation cables and other equipment being installed in their close vicinity, even if they are not directly electrically connected. The magnets are therefore labeled both in their physical and electrical order (see Fig. 1).

The type and direction of the propagation give insight into the mutual interaction of circuit components during a FPA. A distinction is made between the following propagation types:

- Propagation along the electrical position of the circuit: If the weights $\mathbf{H}$ of a spectral component decrease continuously in both directions with the electrical positions, the spectral component propagates through the electrical wiring.
- Propagation along the physical position of the circuit: If the weights $\mathbf{H}$ decrease with the physical position, the spectral component propagates through the helium or the mechanical connection between the magnets. Propagation in helium is usually limited to three physically close magnets, which are installed together in adjacent cryostats belonging to the same cryogenic cell. It is further possible that the instrumentation wires or power cables of nearby physical neighbors cause interference and generate noise, which also propagates along the physical position.
- Artifact of the QDS measurement unit: Lastly, the propagation can depend on effects in the QDS measurement unit. One QDS measurement unit measures the voltage signals on one to three electrically close magnets and one reference magnet. If the QDS measurement unit's input position affects the spectral component, it is assumed that the voltage signal is generated by the measurement unit's electronics, not present directly at the magnet [3], [4].

In addition to the characteristics of the spectral components, several correlation parameters are considered to interpret the spectral components. The most relevant of these are:

- The magnet manufacturer: Each of the three manufacturers used slightly different materials and fabrication methods to produce the magnets, which results in slightly different magnet behavior.
- The sector number: The circuit layout and hardware component manufacturers vary slightly across each of the eight sectors.
- The amplitude and the ramp rate of the circuit current at the time of the FPA trigger: These one relate to the stored energy in the circuit and affect the voltage amplitude after the triggering of the FPA. The ramp rate dI/dt refers to the change of the current over time.
- The FPA event type: The presence of a magnet quench during the analyzed time period affects the frequency spectra significantly. The activation of the QHs induces new voltages, and the diode opening leads to additional transient voltages visible in the frequency spectra.

In the results section, the listed propagation types and correlation parameters are determined and relationships are established for each spectral component to identify the spectral components' underlying potential physical processes.

E. Anomaly Detection

FPA events are abnormal when a frequency spectrum cannot be reconstructed well with the learned spectral components. For this purpose, for each FPA event $q=1,...,Q$ the maximum reconstruction loss $\hat{d}_q$ over all signals in the FPA event is calculated. This reconstruction loss depends on the chosen hyperparameters. Those are the combinations of distance measure $d_*(\cdot)$, the number of spectral components K, and the selected FFT window functions. Hence, anomalies with abnormal behavior are estimated across all hyperparameter combinations.

To make the maximum reconstruction loss $\hat{d}_q$ of different combinations of hyperparameters comparable, the probability

distribution over $\hat{d}_q$ is calculated. The probability distribution over $\hat{d}_q$ for each hyperparameter combination is assumed to be gamma distributed,

$$f(\hat{d}_q;\alpha,\theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}}\hat{d}_q^{\alpha-1}e^{-\frac{\hat{d}}{\theta}}, \tag{9}$$

where Γ is the Gamma function. The parameters α and θ are derived through a maximum likelihood estimation [42] to fit the distribution of the maximum reconstruction losses over all FPA events for one combination of hyperparameters. For this distribution, a p-value z of an event can be defined by the probability of obtaining a $\hat{d}_q'$ at least as high as the observed $\hat{d}_q$,

$$z = \int_{\hat{d}_q'}^{\infty} f(\hat{d}_q;\alpha,\theta)d\hat{d}_q. \tag{10}$$

The distribution fit is performed for all possible combinations of hyperparameters. Therefore, each event has as many p-values as there are combinations of hyperparameters. Abnormal events are then considered as those for which the median of all p-values is low. This yields an anomaly identification strategy that is robust to the choice of hyperparameters.

As mentioned earlier, the 48 events without a quench and the 494 events with a single quench are considered for anomaly detection. Another restriction concerns the selection of distance measures. Here, the IS-divergence is not considered for anomaly detection. Anomalies that indicate critical hardware faults are expected to have dominant amplitudes, but the IS-divergence compares relative amplitudes. The IS-divergence is, however, still relevant for identifying components explaining physical processes behind anomalies. Thus, for anomaly detection, only two distance measures are used: The Euclidean distance and KL-divergence. Together with the seven FFT window functions and the 19 different numbers of spectral components, 266 combinations of hyperparameters are used for anomaly detection.

F. Anomaly Interpretation

Anomalies are interpreted by considering the spectral components' weights in the FPA voltages. If the weights of certain spectral components are higher in a FPA event with high reconstruction loss, these spectral components might be associated with the anomaly. Thus, with the knowledge gained in Section II-D, also the physical process underlying the spectral component can be attributed to the anomaly. This is used to check whether the anomaly is pointing at a hardware problem.

III. Results

A. Spectral Component Selection

Contrary to anomaly detection, all 399 combinations of hyperparameters and all 699 FPA events, are used to derive the spectral components for interpretation. Figure 5 shows the resulting mean pairwise Chebyshev distance $\bar{d}_{Ch}$ of the (a) spectral components $\mathbf{W}$ and their (b) corresponding weights $\mathbf{H}$, as a function of the number of spectral components

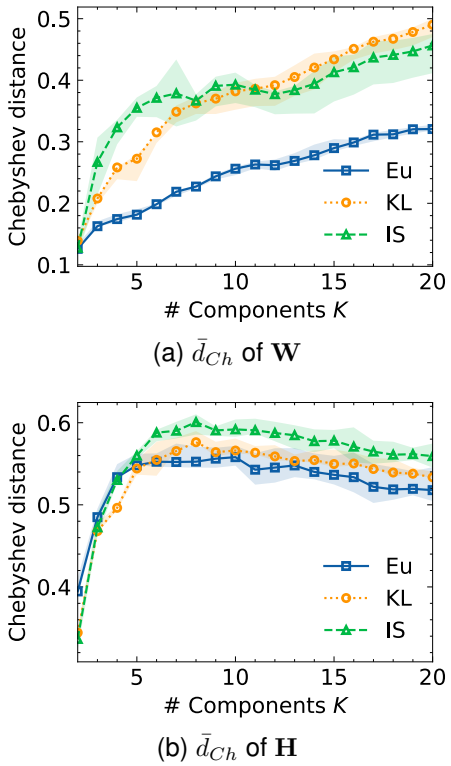


Fig. 5. The mean pairwise Chebyshev distance $\bar{d}_{Ch}$ for (a) the spectral components $\mathbf{W}$ and (b) their weights $\mathbf{H}$ as a function of the number of extracted spectral components K. Compared are the NMF distance measures: The *Squared Euclidean (Eu) distance*, the *Generalized Kullback-Leibler (KL) divergence*, and the *Itakura-Saito (IS) divergence*. The curves indicate the mean over seven different FFT windows, with the first and third quartiles defining the lower and upper confidence intervals, respectively.

K for the three distance measures in Eq. 3-5. The average $\bar{d}_{Ch}$ values across seven distinct FFT windows are shown, with the first and third quartiles forming the lower and upper confidence intervals, respectively. In the range of around seven components, there is a local maximum visible, except for the Eu-distance and KL-divergence in Fig. 5a. Although the $\bar{d}_{Ch}$ curves in Fig. 5a further increase after this extreme point, they decrease in Fig. 5b. The IS-divergence shows the best performance for less than eleven spectral components in Fig. 5a and for more than four spectral components in Fig. 5b. In these regions, the effect of considering relative amplitudes by the IS-divergence is reflected: Also frequencies with small amplitudes are well reconstructed, which increases the diversity of the spectral components and their weights. A similar behaviour is observed for the KL-divergence.

Compared to the number of spectral components or the distance measure, the impact of the FFT window function on $\bar{d}_{Ch}$ is lower, as shown by the relatively narrow confidence intervals in Fig. 5a and 5b. To choose the ideal FFT window function, the FFT window function with the highest $\bar{d}_{Ch}$ at the local maxima of the curves at $K = 7$ is selected. At $K = 7$, the Chebyshev distance is highest if the FFT is calculated using a Hanning window function (not shown in plot). Hence, frequency spectra, derived with a Hanning FFT window function and reconstructed with seven components,

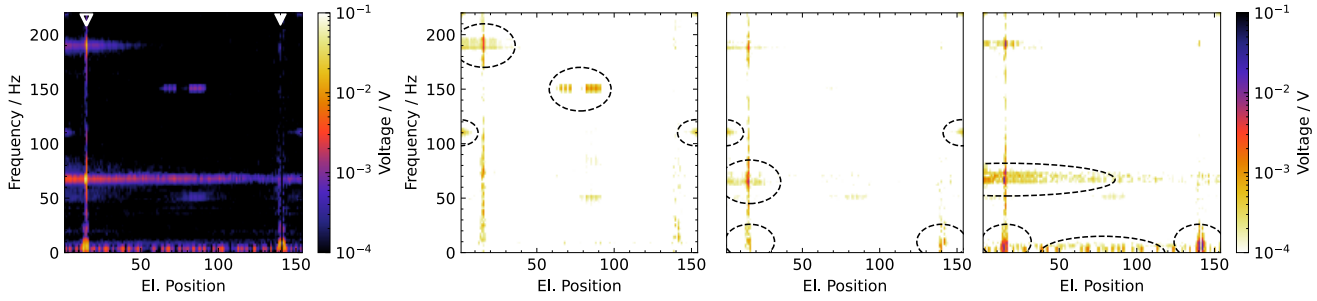


Fig. 6. Comparison of three different distance measures with a FPM. An FPM shows the frequency and amplitude of the voltage signal as a function of the electrical position. (a) shows a FPA event with a quench in sector 78 on 31.03.2021 with an FPM. The solid white arrow marks the quenched magnet, while the empty arrow marks its physical neighbors. This FPA event was reconstructed with $K = 7$ and (b) the Eu-distance, (c) the KL-divergence, and (d) the IS-divergence. To facilitate visual comparison, the values from the input FPA event in (a) are subtracted from those of the reconstructed events. The absolute values of the subtractions are shown as an FPM in (b-d). Circled areas highlight incomplete reconstructions.

are shown in more detail in Fig. 6.

To illustrate the propagation of frequencies in a FPA event, Frequency-Position Maps (FPMs) are used. An example of such an FPM is shown in Fig. 6a, where the frequencies occurring in the 154 voltage signals, measured [0.2; 0.575] seconds after the triggering of the FPA, are plotted as a function of the electrical position for the FPA event on 31.03.2021 in sector 78. The FPM in Fig. 6a shows the processed and Fourier-transformed voltage signals from Fig. 2 as frequency spectra values $v_{i,j}$. During this event, a quench occurred at the electrical position 141 (white solid arrow in Fig. 6a). The electrical positions 14 and 15 are the physical circuit neighbors of the quenched magnet (white empty arrow in Fig. 6a).

Figures 6b-6d, show the absolute difference $|v_{i,j} - \hat{v}_{i,j,*}|$ between the NMF reconstructions $\hat{v}_{i,j,*}$ and the input $v_{i,j}$ for the different distance measures discussed above. Only frequencies below 220 Hz are included for better visibility. The example aims at comparing the reconstructions for Eu-distance, KL-divergence, and IS-divergence using $K = 7$ and a Hanning FFT window function. The colored spots in the FPMs (Figs. 6b-6d) show electrical positions and frequencies with a reconstruction difference. The darker the color of the points, the larger is the reconstruction difference. If the reconstruction is identical to the input and the reconstruction difference is zero, the plots would be completely white.

In Fig. 6b, small reconstruction differences are visible in the low-frequency range for the Eu-distance. Voltage amplitudes in this range are generally higher. Reconstructions optimized with the Eu-distance, demonstrate superior performance in reconstructing lower frequencies as compared to the two other measures. At 110 Hz, 150 Hz, and 180 Hz significant reconstruction differences are visible by dark spots, highlighted by dashed ellipses. At these frequencies, the amplitudes are lower and are therefore not taken into account by the Eu-distance.

In comparison, for the KL-divergence more significant reconstruction differences are visible by the dark spots in the low-frequency range in Fig. 6c. Instead, the reconstruction differences at 150 Hz and 180 Hz are smaller than for the Eu-distance measure. This can be explained by the fact that the KL-divergence reflects the relative entropy. Hence, instead of reconstructing the low frequencies with high amplitude more accurately, the entropy is optimized by reconstructing frequencies with lower amplitudes as well.

In Fig. 6d, the IS-divergence leads to reconstructions where both 110 Hz and 150 Hz oscillations have been captured, as there are no significant reconstruction differences visible there. The scale-invariance [34] of the IS-divergence makes it ideal to reproduce also frequencies with low amplitude. However, the low-frequency range, where significant reconstruction differences are visible, has been reconstructed poorly by the IS-divergence.

Based on the interpretation of Fig. 6 and further visual inspections, four Eu-distance components and three IS-divergence components that capture the frequency spectra best were selected for further analysis. Figure 7 shows how the selected spectral components are used to reconstruct the frequency spectra in the voltage signal, measured [0.2; 0.575] seconds after the triggering of the FPA event. Figures 7b-7h show the contribution $\hat{v}_{i,j}$ of each of the seven selected spectral components j to the reconstruction of the frequency spectra $v_{i,j}$ of the FPA event in Fig. 7a. In all FPMs, the frequencies and amplitudes of the voltage signals are shown as a function of the electrical position. The amplitude is displayed logarithmically as a color in the range 10^{-4} V to 10^{x} V, where x is the maximum amplitude of the spectral component in this event. This x is indicated in the caption of each FMP with the spectral component number j. For the reconstruction of the sub-spectra with the different components, the following can be observed: High amplitudes in the low-frequency range, with their maxima at (b) 3 Hz, (c) 6 Hz, (d) 20 Hz, and (e) 66 Hz, are reconstructed by the Eu-distance spectral components. Lower amplitudes in the high-frequency range are reconstructed with the IS-distance spectral components, having their maxima at (f) 150 Hz and (g) 478 Hz. Lastly, a broadband spectrum, spanning vertically over the whole frequency range, can be reconstructed by the IS-distance spectral component (h). Due to the additive nature of NMF the Eu-distance reconstruction in Figs. 7b-7e and the

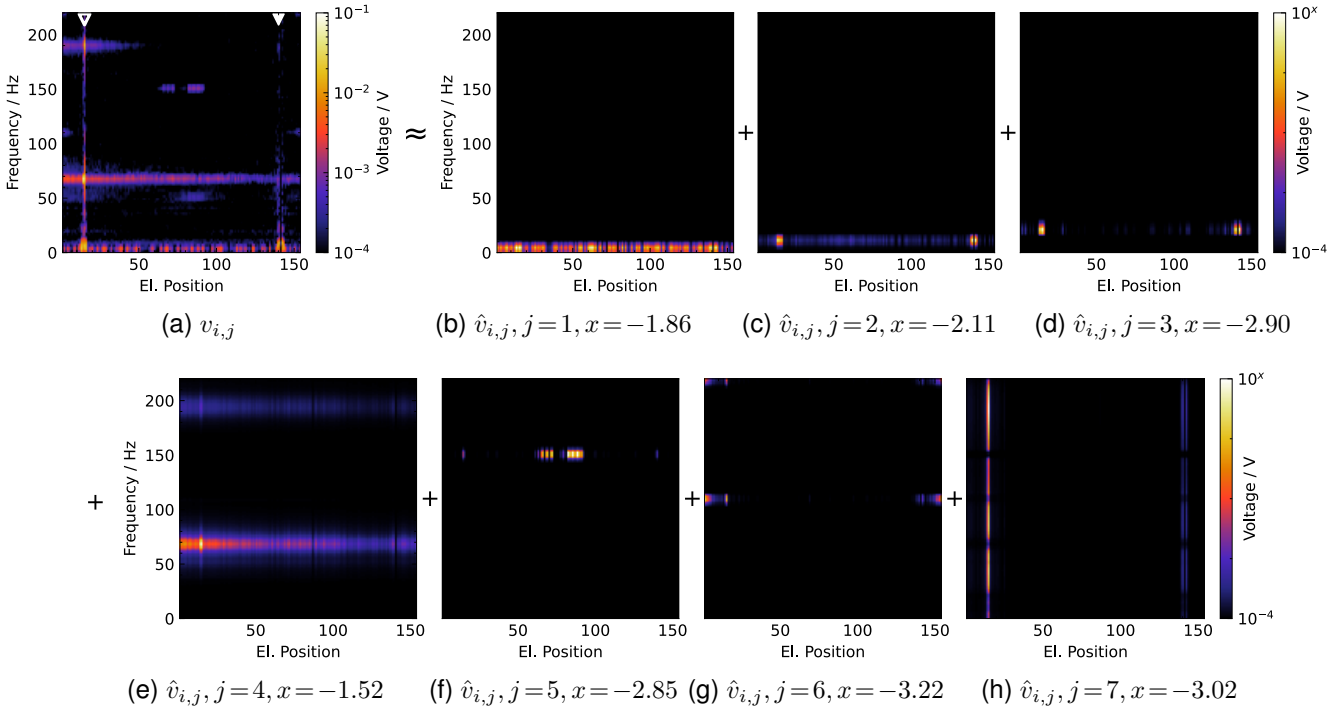


Fig. 7. Frequency and amplitude of the identified seven spectral components as a function of the electrical position. (a) shows the FPM of the frequencies occurring in the voltage signal, measured [0.2; 0.575] seconds after the triggering of the FPA of sector 78 on 31.03.2021. (b-h) show $\hat{v}_{i,j}$ for each spectral component j, used to reconstruct the initial FPM in (a). For better visibility, the maximum of the color axis is scaled with 10^x V. Additionally, the frequency range is restricted to 0-220 Hz, in which the majority of the spectral components occur.

IS-divergence reconstructions in Figs. 7f-7h can be added to reconstruct the input in Fig. 7a. The propagation and physical process of each spectral component are discussed in the next subsection.

B. Spectral Component Interpretation

Seven spectral components have been identified and are considered important to describe the overall frequency response of the LHC's main dipole circuits during FPAs. In the following, their characteristics and the potential underlying physical processes are discussed one by one. A summary of the discussed spectral components is given in Table I, where the columns show the characteristics described in Section II-D.

- Spectral component one (SC1) is visible in Fig. 7b in the bright horizontal frequency band at 3 Hz. There are particularly bright spots at positions 15 and 141 which have a different explanation than the remaining bright spots.
  The magnets at positions 15 and 141 are the physical and electrical neighbors of the quenched magnet, respectively. Considering all FPA events with a quench, the average maximum amplitude at the physical and electrical neighbors of the quenched magnet is 62 mV. In FPA events without a quench, the average maximal amplitude is 5 mV. It can be concluded that the physical process causing SC1 is the quench of a magnet. It can be assumed that the quenching magnet causes electromagnetic perturbations, which are propagating through instrumentation wires and the connected QDS measurement units. Interestingly, the 3 Hz frequency amplitude is one order of magnitude larger if the quenched magnet was produced by manufacturer one. These perturbations are important because they are most likely the origin of high-energy secondary quenches in neighboring magnets tens of milliseconds after the primary quench occurred.
  The remaining bright spots are introduced in the pre-processing steps. Deviations of the signal $\mathbf{x}$ from the exponential trend $\bar{\mathbf{x}}$ (see Eq. 1), are interpreted as oscillation by the FFT. These oscillations are part of SC1 but do not originate from a physical process.
- Spectral component two (SC2) is visible in Fig. 7c by two bright points at 6 Hz at the electrical positions around 15 and 141.
  These are the locations of the physical and electrical neighbors of the quenched magnet. As for SC1, it can be assumed that the physical processes causing SC2 are electromagnetic perturbations induced by the quenched magnet. This assumption is supported by the fact that the average maximum voltage of FPA events is 36 mV, while for events without quench the average voltage is 100 µ V. Similarly to SC1, the frequency amplitudes are one order of magnitude larger if the quenched magnet was produced by manufacturer one.
- Spectral component three (SC3), in Fig. 7d, shows a similar pattern to SC2. Bright spots are visible at the physical and electrical neighbors of the quenched

TABLE I
CHARACTERISTICS OF SPECTRAL COMPONENTS.

Spectral Component	Dominant Frequencies	Location of Maximum	Average Maximum Amplitude	Propagation	Possible Physical Process
SC1 Fig. 7b	3 Hz	Phys. & el. neighbors of quenched magnet	62 mV	Phys. & el. position	Electromagnetic perturbation
		Evenly distributed within circuit	5 mV	Evenly distributed within circuit	Preprocessing
SC2 Fig. 7c	6 Hz	Phys. & el. neighbors of quenched magnet	36 mV	Phys. & el. position	Electromagnetic perturbation
SC3 Fig. 7d	20 Hz	Constant across all magnets	14 mV	Constant across all magnets	Diode induced oscillation
		Phys. & el. neighbors of quenched magnet	4 mV	Phys. & el. position	Electromagnetic perturbation
		Phys. & el. neighbors of power converter	1 mV	Phys. & el. position	Leftover voltage waves traveling along the chain of magnets by magnet impedance
SC4 Fig. 7e	66 Hz 184 Hz 302 Hz	Phys. neighbors of quenched magnet	73 mV	El. position	Oscillations caused by quench
SC5 Fig. 7f	150 Hz	El. neighbors of EE systems	1 mV	Position in QDS measurement unit	Artifact of the QDS measurement unit
SC6 Fig. 7g	107 Hz 220 Hz 260 Hz 370 Hz 478 Hz	Phys. & el. neighbors of PC	690 µV	El. position	Passive hardware elements of PC in sector 78
SC7 Fig. 7h	Broadband spectrum	Phys. & el. neighbors of quenched magnet	4 mV	Phys. & el. position	Quench heater induced oscillation
		Phys. & el. neighbors of quenched magnet	1 mV	Phys. position	Quench dependent oscillations

magnet. However, additional physical processes appear when comparing different events where either no quench occurs, a single quench occurs, or a diode opened due to a secondary quench in the analyzed time window.

In events with an additional diode opening during an EE plateau, the average maximum amplitudes are 14 mV, where the values show little variance. Here, the diode opening induces a 20 Hz oscillation in the circuit, which is constant across all magnets. No diode opened during the FPA event shown in Fig. 7d, which is why the bright spots are not visible across all electrical positions.

In events with a single quench the average maximum amplitudes are highest at the physical and electrical neighbors, with 4 mV. This effect can be seen in Fig. 7d. Similar to SC1 and SC2, the physical process causing SC3 is electromagnetic perturbations originated at by the quenched magnet.

In events with no quenches the amplitude of SC3 is highest at the magnet close to the PC with an average maximum amplitude of 1 mV. Here, the amplitude gradually decreases and is lowest at the first EE system. This can be traced back to the quench-independent leftover voltage waves traveling along the chain of magnets as governed by the magnet impedance. This effect is observed for all events and is proportional to the amplitude and the ramp rate of the circuit current at the moment of the FPA trigger [17]. In Fig. 7d the frequencies caused by the quench are more prominent, making this process not observable with the given color range.

- Spectral component four (SC4) is visible in Fig. 7e and shows a bright spot at 66 Hz at positions 14 and 15. This shows that the locations of its voltage maxima are in the physical neighbors of the quenched magnet. From there the bright spot is gradually getting darker in both directions, indicating that the oscillation is propagating along the electrical direction.

  A similar pattern is observed at 184 Hz, but in a darker color. In addition, SC4 is high at 302 Hz. The amplitude of this approximate 3^{rd} and 5^{th} harmonic scales indirectly proportional to the number of the n^{th} harmonic.

  While the exact physical process of SC4 remains elusive, it is expected that it is emphasized by a quench. This expectation is supported by comparing FPA events with and without a quench. In events without a quench, the average maximum amplitude is two orders of magnitude lower than for events with a quench (73 mV vs. 730 µV).
- Spectral component five (SC5) appears as a double horizontal band at 150 Hz near the center around the electrical position 77 in Fig. 7f. The bright spots of the band occur at exactly the same input of each QDS measurement unit. This indicates that SC5 is introduced by the electronics of the QDS measurement unit. SC5 only occurs in FPA events in sectors 12, 45, 67, 78, and 81. No specific hardware component has been identified as the cause of this behavior. It occurs in the same

position, regardless of the number of quenches or the quench position, with an average maximum amplitude of 1 mV.

- Spectral component six (SC6) is visible at 107 Hz and 220 Hz as spots originating at the electrical positions 1 and 154 of Fig. 7g. The magnets at these electrical positions are installed close to the power converters of the circuit. The amplitude of SC6 decreases with increasing distance from the power converter. In addition, SC6 has high amplitudes at 260 Hz, 370 Hz, and 478 Hz, not visible in Fig. 7g due to the restricted range of the frequency axis.
  During the EE plateaus the PC is deactivated, indicating that SC6 originates from passive hardware components in the PC. SC6 only occurs in sector 78. No exact hardware component has been identified that could explain this behavior. It always occurs in the same electrical positions, regardless of the quench position, with an average maximum amplitude of 690 µV.
- Spectral component seven (SC7) in Fig. 7h shows one vertical line with high amplitude at the electrical positions 14 and 15, and one line with low amplitude at the electrical positions around 140 and 142. Both lines have interruptions at frequencies already reconstructed by other spectral components. These vertical lines indicate a broadband spectrum in magnets physically close to the quench. In the time domain, this broadband spectrum corresponds to a spike. In a previous analysis of faults in a subsystem of the LHC protection system, such spikes were used as indicators for intermittent short circuits [19]. Hence, SC7 might also be a critical indicator of an intermittent short circuit in the magnet.
  In FPA events with a single quench the average maximum amplitude of this broadband spectrum is 1 mV. For FPA events without quenches, the average maximum amplitude of SC7 is significantly smaller ($<$100 µV). This shows that SC7 depends on the quench.
  A different physical process is observed during events with secondary quenches, where the average maximum amplitude is 4 mV if the QHs of one of the additionally quenched magnets are activated during the EE plateaus. In this case, SC7 propagates along the physical and electrical position of all magnets in the circuit and is likely induced by the inductive part of the QH strips. No QHs were activated during the FPA event shown in Fig. 7d, which is why this physical process cannot be observed there.

C. Anomaly Detection

In this subsection, the selected anomalies from FPA events with low median p-values following the definitions given in Section II-E are presented. FPA events with low median p-values are labeled based on an LHC specific four-digit identifier of the quenched magnet, marked by a '#'.

Figure 8 shows a boxplot, a conventional method to illustrate statistical data characteristics [43], of the ten FPA events with the lowest median p-value. For each anomalous FPA event, the quenched magnet is given on the x-axis and a box represents the statistical distribution of p-values over 266 different combinations of hyperparameters (the IS-divergence is not used). The box represents the range between the first and the third quartile, where the line in the middle represents the median. The outer limits further indicate the variability of p-values, which are obtained by subtracting 1.5 times the box length from the first quartile and adding it to the third quartile, respectively. The y-axis is plotted logarithmically, therefore, the first quartile's outer limit is cut, if it is zero.

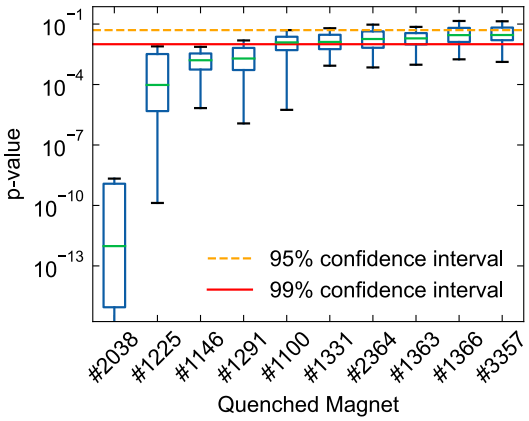


Fig. 8. Boxplot of the 10 FPA events with the smallest median p-values. For each event, the quenched magnet is given on the x-axis. Each box extends from the first to the third quartile, with a horizontal line at the median. The lines extending from the box further show the variability of p-values. They cover a range of 1.5 times the box length, either subtracted from the first quartile or added to the third quartile. The orange dashed line indicates the confidence interval of 95% at a p-value of 0.05, while the red vertical line indicates the 99% confidence interval at a p-value of 0.01.

The four FPA events with the quenched magnets #2038, #1225, #1146, and #1291 state a median p-value smaller than 1%. In the boxplot for those FPA events, the first and third quartiles are below the 99% confidence interval, and the outer box limits are below 95%. For the six other FPA events in Fig. 8, both quartiles and outer limits are above their respective 99% and 95% intervals. Based on this classification, only the four events with a median p-value of less than 1% are therefore referred to as anomalies. In the following subsection, these anomalies are discussed in more detail.

D. Anomaly Interpretation

To better understand the characteristics of the anomalies, the spectral component weights in the anomalous FPA events are compared to those in normal FPA events. Notably, the weights of SC7 are elevated in three of the four identified FPA events. As discussed above, SC7 represents a broadband spectrum and has an average maximum amplitude of 1 mV for FPA events with a single quench (see Tab. I). However, with 820 µV only the maximum amplitude in the anomalous FPA event where the magnet #1146 quenched is similar to this average. For the FPA events where the magnets #2038, #1225, and #1291 quenched the amplitudes of this spectral component are 240 mV, 80 mV, and 210 mV, respectively. These amplitudes are more than 80 times higher than the

maximum average amplitude for other FPA events. In no other FPA event with a single quench the values are this high. It is inferred that a high SC7 in the quenched magnet is a sufficient criterion for identifying an anomaly.

Having identified the criticality of a high SC7 amplitude in the quenched magnet, the previously excluded 157 FPA events with secondary quenches are also examined for this characteristic. One FPA event with an amplitude of 1200 mV, more than 1200 times higher than the average maximum, stands out. Therefore, this FPA event, where a quench occurred at the magnet #2421, is also referred to as an anomaly in the course of this analysis.

In magnet #1146, the SC7 amplitude is not elevated during the quench. Instead, the low p-value in this FPA event is associated with a high SC1 amplitude of 1500 mV in the physical neighbors of the quenched magnet. As discussed in Subsection III-B, this component represents electromagnetic perturbations propagating through instrumentation wires and the connected QDS measurement units. Based on current knowledge, these electromagnetic perturbations are not associated with a hardware fault.

E. Recommended Maintenance Actions

One of the four quenched magnets, with a significantly increased SC7 amplitude during a FPA event, has developed an intermittent short circuit during the FPA event on 25.04.2021. As such an intermittent short circuit is a critical event, the other three magnets #1225, #1291, and #2421 are also treated as potentially critical and will be checked by transient voltage measurement. If an intermittent short circuit cannot be excluded during the transient voltage measurement, these magnets could be replaced in one of the next maintenance stops of the LHC. In any case, the electronics of the QDS measurement units of these magnets should be exchanged in order to exclude measurement errors.

Transient measurements will also be performed on the magnet #1146. These measurements will provide further information about electromagnetic perturbations.

Tab. II summarizes the five discussed anomalies, sorted by their median p-value in the first column. The second column shows the quenched magnet, the affected circuit and the date of the related FPA event. The third column summarizes main findings of the FPA event and states the recommended maintenance actions.

TABLE II
LIST OF DETECTED ANOMALIES WITH RECOMMENDED MAINTENANCE ACTIONS IN THE REMARKS COLUMN.

Median p-value	FPA Event	Remarks
8×10^{-11}	#2038 Sector 78 25.04.2021	High SC7 in #2038 (240 mV) Exchanged on 25.04.2021 due to intermittent short circuit
7×10^{-5}	#1225 Sector 45 12.05.2021	High SC7 in #1225 (80 mV) Additional measurements Hardware replacements
1×10^{-3}	#1146 Sector 34 06.05.2021	High SC1 (1500 mV) Additional measurements
2×10^{-3}	#1291 Sector 12 14.05.2021	High SC7 in #1291 (210 mV) Additional measurements Hardware replacements
-	#2421 Sector 34 20.04.2021	High SC7 in #2421 (1200 mV) Additional measurements Hardware replacements

IV. CONCLUSION

In this study, the voltages measured across the 1232 magnets in the eight LHC main dipole circuits were analyzed to understand the normal and abnormal behavior of the circuits. Specifically, the amplitude and propagation of the frequency spectra measured at the magnets in 699 Fast Power Abort (FPA) events were investigated using Non-negative Matrix Factorization (NMF).

This allowed the extraction of seven spectral components that define normal behavior, occurring in the measured voltages during a FPA event. Analyzing the spectral components' distribution and propagation across the circuit and across FPA events provided a deeper understanding of the mutual interaction of hardware components and allowed identifying the potential physical processes causing the spectral components. It was shown that spectral components one to three, with maxima at 3 Hz, 6 Hz, 20 Hz, are induced by the quench due to electromagnetic perturbations. Their amplitudes are one order of magnitude higher when the quenched magnet was produced by manufacturer one. Spectral component four shows a 66 Hz oscillation induced by the quench. With maxima at 150 Hz and 478 Hz, components five and six are independent of the quench and were attributed to artifacts of the QDS measurement unit, and to passive elements in the power converter of one individual circuit. Spectral Component seven (SC7) shows a broadband spectrum, induced by the quench. As previous studies showed that such broadband spectra can be an indicator of short circuits [19], SC7 could also indicate an intermittent short circuit in the magnet.

Five magnets with abnormal behavior during FPA events were detected using the reconstruction loss of NMF and the SC7 amplitude at the quenched magnet. One of these magnets was replaced on 25.04.2021 after a short circuit was detected following the FPA event. Similarly to the replaced magnet, three of the four remaining magnets showed an elevated SC7 amplitude during their quench, which is more than 80 times higher than normal. Dedicated transient measurements will be performed on these magnets and the electronics of their QDS measurement unit should be replaced. If an intermittent short circuit still cannot be excluded, the three magnets could be replaced in one of the next maintenance stops of the LHC to prevent weeks of unplanned LHC downtime. In the magnet which did not show a high SC7 during the FPA event in the quenched magnet, data do not indicate a hardware fault. Instead, an abnormally high spectral component one was observed, which will also be evaluated by transient measurements. The presented methodology has proven to be a powerful tool to describe the normal behavior of the circuits systematically and to detect abnormal behavior indicating potential hardware fatigue and degradation of hardware components in the circuit.

REFERENCES

[1] J. Schultz, "Protection of Superconducting Magnets," *IEEE Transactions on Applied Superconductivity*, vol. 12, no. 1, pp. 1390–1395, 2002.

[2] E. Ravaioli, K. Dahlerup-Petersen, F. Formenti, V. Montabonnet, M. Pojer, R. Schmidt, A. Siemko, M. Solfaroli Camillocci, J. Steckert, H. Thiesen, and A. Verweij, "Impact of the Voltage Transients After a Fast Power Abort on the Quench Detection System in the LHC Main Dipole Chain," *IEEE Transactions on Applied Superconductivity*, vol. 22, no. 3, pp. 9002504-9002504, 2012.

[3] R. Denz and F. Rodriguez-Mateos, "Electronic Systems for the Protection of Superconducting Elements in the LHC," *IEEE Transactions on Applied Superconductivity*, vol. 16, no. 2, pp. 1725-1728, 2006.

[4] R. Denz, K. Dahlerup-Petersen, F. Formenti, K. H. Meß, A. Siemko, J. Steckert, and L. Walckiers, "Upgrade of the Protection System for Superconducting Circuits in the LHC," *Proceedings of the Particle Accelerator Conference*, vol. 23, pp. 244-246, 2009.

[5] F. Rodriguez-Mateos, P. Pugnat, S. Sanfilippo, R. Schmidt, A. Siemko, and F. Sonnemann, "Quench Heater Experiments on the LHC Main Superconducting Magnets," *Proceedings of the Particle Accelerator Conference*, vol. 1-3, pp. 2154-2156, 2000.

[6] F. Sonnemann, "Resistive Transition and Protection of LHC Superconducting Cables and Magnets," Ph.D. thesis, RWTH Aachen University, 2001.

[7] F. Bourgeois and K. Dahlerup-Petersen, "Methods and Results of Modeling and Transmission Line Calculations of the Superconducting Dipole Chains of CERN's LHC Collider," *LHC Project Report 497*, CERN, 2001.

[8] K. Dahlerup-Petersen, F. Rodriguez-Mateos, R. Schmidt, F. Sonnemann, "Energy Extraction for the LHC Superconducting Circuits," *Proceedings of the Particle Accelerator Conference*, Vol. 5, pp. 3448-3450, 2002.

[9] E. Ravaioli, K. Dahlerup-Petersen, F. Formenti, J. Steckert, H. Thiesen, and A. Verweij, "Modeling of the Voltage Waves in the LHC Main Dipole Circuits," *IEEE Transactions on Applied Superconductivity*, vol. 22, no. 3, pp. 9002704-9002704, 2012.

[10] A. Verweij, V. Baggiolini, A. Ballarino, B. Bellesia, F. Bordry, A. Cantone, M. Casas Lino, A. Serra, C. Trello, N. Catalan Lasheras, Z. Charifoulline, G. Coelingh, K. Dahlerup-Petersen, G. D'Angelo, R. Denz, S. Feher, R. Flora, M. Gruwé, V. Kain, and M. Zerlauth, "Performance of the Main Dipole Magnet Circuits of the LHC during Commissioning," *Proceedings of European Particle Accelerator Conference*, vol. 11, pp. 2473–2475, 2008.

[11] L. Coull, D. Hagedorn, G. Krainz, F. Rodriguez-Mateos, R. Schmidt, et al., "Quench Propagation Tests on the LHC Superconducting Magnet String," *LHC Tech. Rep.*, no. 70, 1996.

[12] A. Siemko, "Magnet Quench Process," *CERN Tech. Rep.*, no. 567209, 2001.

[13] R. Saederup, "Local Transfer Function Measurement Data Analysis," *CERN Tech. Rep.*, no. 2675917, 2021.

[14] E. Ravaioli, A. P. Verweij, and H. H. J. ten Kate, "Unbalanced Impedance of the Aperture Coils of Some LHC Main Dipole Magnets," *IEEE Transactions on Applied Superconductivity*, vol. 23, no. 3, pp. 4000104-4000104, 2013.

[15] E. Ravaioli, "First Analyses and Conclusions on the Fast Power Abort Measurements in Sector 67 During the Long Shut-Down 2010-11," *CERN Tech. Rep.*, no. 1137620, 2011.

[16] STEAM website, https://espace.cern.ch/steam/ (accessed Oct. 16, 2023).

[17] M. Janitschke, "Framework for automatic superconducting magnet model generation & validation against transients measured in LHC magnets," Master's thesis, Technical University of Berlin, 2021.

[18] D. Hoang, C. Boffo, N. Tran, S. Krave, S. Kazi, S. Stoynev, and V. Marinozzi, "Intelliquench: an adaptive machine learning system for detection of superconducting magnet quenches," *IEEE Transactions on Applied Superconductivity*, vol. 31, no. 5, pp. 1–5, 2021.

[19] Z. Charifoulline, L. Bortot, R. Denz, F. Rodriguez Mateos, A. Siemko, J. Steckert, A. Verweij, and G. Willering, "Overview of the performance of quench heaters for high-current LHC superconducting magnets," *IEEE Transactions on Applied Superconductivity*, vol. 27, no. 4, pp. 1–5, 2016.

[20] M. Wielgosz, A. Skoczeń, and M. Mertik, "Using LSTM recurrent neural networks for monitoring the LHC superconducting magnets," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 867, no. 1, pp. 40–50, 2017.

[21] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," *AI Magazine*, vol. 38, no. 3, pp. 50-57, 2017.

[22] J.M. Be'itez, J.L. Castro, and I. Requena, "Are artificial neural networks black boxes?," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1156-1164, 1997.

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.

[24] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[25] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971-982, 2013.

[26] M. R. Blanton and S. Roweis, "K-corrections and filter transformations in the ultraviolet, optical, and near-infrared," *The Astronomical Journal*, vol. 133, no. 2, pp. 734-754, 2007.

[27] H. J. Nussbaumer, *The Fast Fourier Transform*. Springer, 1981.

[28] A. V. Oppenheim, *Discrete-time Signal Processing*, 1999.

[29] Fredric J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[30] Å. Björck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.

[31] D. Lee and H. Sebastian Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems*, vol. 13, MIT Press, 2000.

[32] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780-791, 2007.

[33] Y. Cao, P.P.B. Eggermont, and S. Terebey, "Cross Burg entropy maximization and its application to ringing suppression in image reconstruction," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 286-292, 1999.

[34] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793-830, 2009.

[35] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350-1362, 2008.

[36] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, 2012.

[37] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with $\ell 0$-constraints," *Neurocomputing*, vol. 80, no. 1, pp. 38-46, 2012.

[38] Cyrus D. Cantrell, *Modern Mathematical Methods for Physicists and Engineers*. Cambridge University Press, 2000. ISBN 0-521-59827-3.

[39] T. Smets, N. Verbeeck, M. Claesen, A. Asperger, G. Griffioen, T. Tousseyn, W. Waelput, E. Waelkens, and B. De Moor, "Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data," *Analytical Chemistry*, vol. 91, no. 9, pp. 5706-5714, 2019.

[40] M. Suresh and I. R. Varughese, "Effect of Similarity Measures on Repetitive Audio Source Separation," *International Journal of Scientific Engineering and Technology Research*, vol. 3, no. 30, pp. 5933-5934, 2014.

[41] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[42] R. J. Rossi, *Mathematical Statistics: An Introduction to Likelihood Based Inference*, John Wiley & Sons, 2018.

[43] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.