

# A Dataset for Fault Classification in Rock Drills, a Fast Oscillating Hydraulic System

Erik Jakobsson<sup>1,2</sup>, Erik Frisk<sup>2</sup>, Robert Pettersson<sup>1</sup>, and Mattias Krysander<sup>2</sup>

<sup>1</sup> *Epiroc Rock Drills AB, Örebro, 702 25, Sweden*  
*erik.jakobsson@epiroc.com*  
*robert.pettersson@epiroc.com*

<sup>2</sup> *Linköping University, Linköping, 581 83, Sweden*  
*erik.frisk@liu.se*  
*mattias.krysander@liu.se*

## ABSTRACT

This work describes the collection and properties of a publicly available rock drill fault classification data set, used for the 2022 PHM Conference Data Challenge. The data is collected from a carefully instrumented hydraulic rock drill, operating in a test cell while inducing a number of faults. Hydraulic pressures are measured at 50kHz at three different locations, resulting in a detailed pressure signature for each fault. Due to wave propagation phenomena, the system is sensitive to individual differences between different rock drills, drills rigs and configurations. Such differences are introduced in the data by altering certain parameters in the test setup. An important part of the data is therefore the availability of No-fault reference cycles, which are supplied for all individuals. These reference cycles give information on how individuals differ from each other, and can be used to improve classification.

## 1. INTRODUCTION

Hydraulic rock drills are used in a wide range of applications where holes are needed in hard rock materials. Such systems, as seen in Figure 1, often operate under high performance demands in harsh environments, with vibrations and moisture. Normally no high resolution measurement data is available from such machines, and especially not in combination with known internal faults. The data set described in this work changes that, and provides normalized pressure measurements from a hydraulic rock drill, operating in a close-to-reality type of test cell, while different faults and other variations are introduced to the system in a controlled manner.

Erik Jakobsson et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



Figure 1. Two rock drills on a drill rig in their natural underground habitat. Asset: Epiroc.

The intended usage for this data set is for the development of time series classification techniques, based on high sampling rate sensor data.

An important property of this type of data is the large influence of wave propagation on the pressure measurements. This causes small external changes that aren't faults, such as changes in hose lengths, to significantly alter the behavior of the pressure signals. To capture such differences, different *individuals* are part of the data. To introduce such individual differences typically requires changing a number of parts physically in the test setup. The number of combinations required, together with the different faults induced, quickly add up to a large number of disassembly/assembly operations. It is also difficult to foresee all differences that may occur for combinations of part tolerances and drill rig configurations in the field. To overcome these problems and to reduce manual

labor, control parameters are used as replacement for physical changes to emulate individual differences.

The differences between some fault classes are small compared to the individual differences, and this is a challenging classification task. Figure 2 shows an example of how a pressure trace in a fault scenario differs from the No-fault case. The oscillation seen around  $t = 9$  ms arrives later as the fault is introduced, but this difference can be masked by individual differences. A known reference from the current individual might be required for calibration.

The data is intended to be used in the following way:

1. Training data is supplied from a number of individuals. Reference data describing the No-fault case is also available.
2. Models are trained using the data.
3. Unseen data is presented from new individuals. Together with this unseen, possibly faulty data, there are a comparably small amount of reference measurements from the No-fault class from each specific individual.
4. The target is to classify the unseen data.

That is, the challenge is to create models able to generalize over such individuals, while having access to some reference data from the specific individual. Hence, it is important to use different individuals in training and validation. The reference data from the validation individual can be considered unknown until a fictive deployment of the method. Hence this validation data can not be used during training.

## 2. APPLICATION

A hydraulic rock drill, as seen mounted on an underground drill rig in Figure 1, is a hydro-mechanical device using for generating stress waves in a drill steel. The rock drill is attached on a sliding cradle at the rear of a beam on the rig. The drill rig positions the beam and supplies the rock drill with hydraulic oil, flushing fluid, feed force and compressed air.

The generated stress waves are transferred through a drill rod to a drill bit where tungsten carbide button bits crushes rock and creates a hole, typically around 30-60 mm in diameter for this type of machine. The drill bit is caused to rotate to allow new uncrushed rock in front of the button bits before the next impact occurs. Some flushing media, air or water, is flushed through the drill steel and bit to remove the debris.

The rock drill consists of four main systems. The percussion system, where an impact piston is caused to oscillate by use of hydraulic fluid power. The rotation system, where a hydraulic motor rotates the drill steel. The damper system, where stress wave reflections from the rock are dissipated as heat, and the flushing system, where flushing media enters the drill steel. This work primarily targets the percussion- and damping sys-

Table 1. Available pressure sensor names and description.

| Sensor | Sampling | Description                               |
|--------|----------|---|
| pin    | 50kHz    | Percussion pressure at inlet fitting.     |
| pdin   | 50kHz    | Damper pressure inside the outer chamber. |
| po     | 50kHz    | Pressure in the volume behind the piston. |

tem. A schematic image of these two systems are seen in Figure 3.

## 3. DATA COLLECTION

A picture from the actual test setup is seen in Figure 4. Pressure measurements on three different locations are listed in Table 1, however the specific model and make of sensors and data acquisition equipment is left undisclosed due to proprietary reasons. The test cell control system allows for automatic test sequences, including set levels of pressures, forces and flows. This functionality is used to emulate the individual differences expected to be found in a large population of rock drills and drill rigs.

The data set contains 11 different classes, including the No-fault class which also serves as reference. The approximate location in the rock drill where faults are introduced are shown in red capital letters in Figure 3. The different classes are listed in Table 2 where the *Letter* column corresponds to the fault location seen in Figure 3. The *Label* corresponds to the class labels used in the first column of the supplied data files.

The automated collection sequence is shown in Figure 5, where two variables P1 and P2 are changed according to the graph, in the same order for each class. The times of collection for the individuals are shown by the numbered intervals.

Most of the induced faults requires disassembly of the rock drill, such as removing seals, orifices and exchanging parts. The following sequence is used for each fault:

1. Disassemble the rock drill.
2. Change / remove / modify parts to induce the different faults.
3. Assemble rock drill.
4. Run test cycle according to Figure 5. The variation of parameters in the cycle gives data for different individuals.
5. Save data for the different individuals.

The structure of the data is shown in Figure 6. In total 8 individuals x 11 classes x 300 to 700 cycles each give approximately 54 000 cycles in total, each with data from 3 sensors. Each pressure time series range from 556 to 748 samples. The size of the dataset in .csv format is 976 MB.

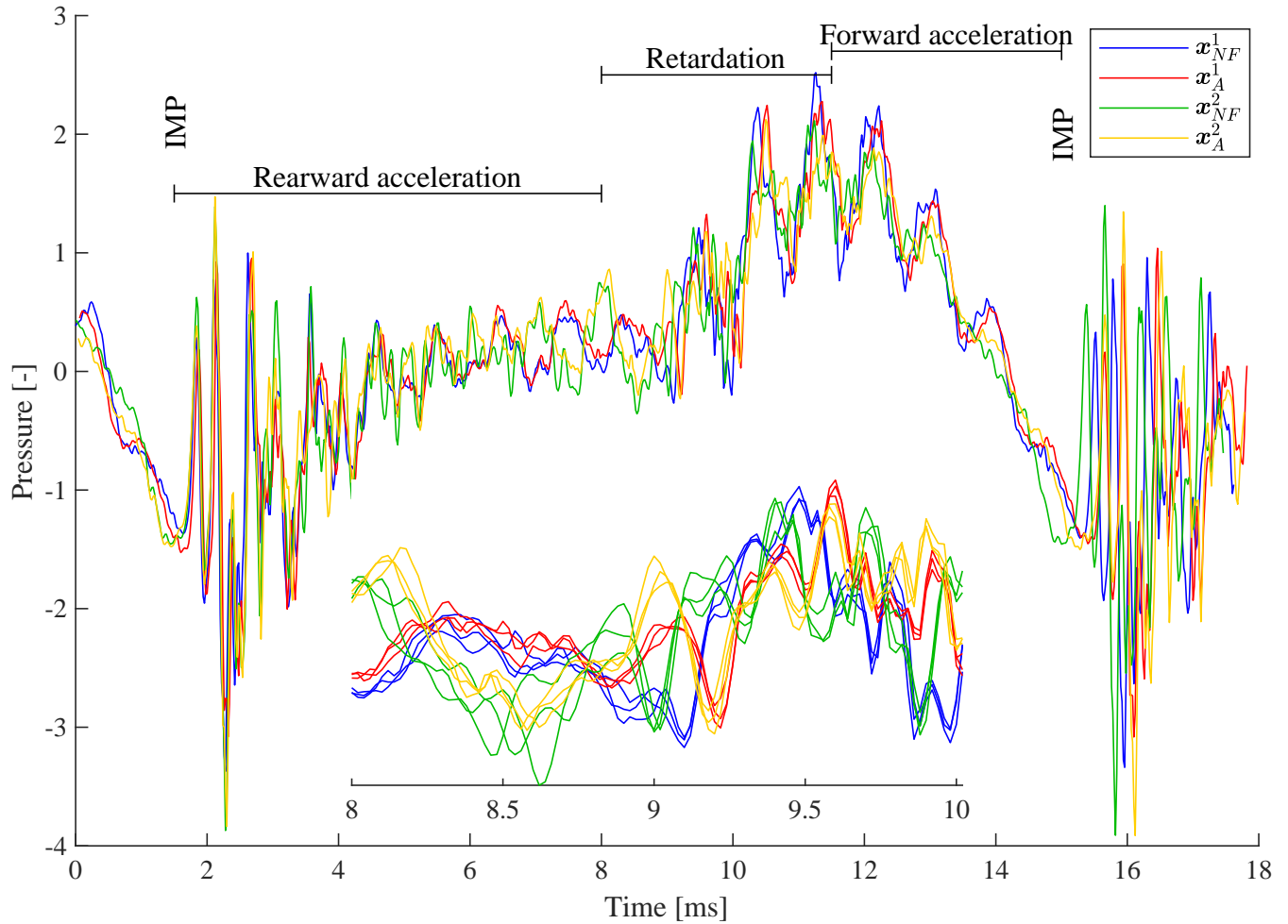


Figure 2. Time series  $x$  from two classes, subscript NF and A, and two individuals, superscript 1 and 2. Multiple cycles are shown to highlight the deterministic properties of the data, i.e., that in a fixed mode and individual, the behavior of the pressure signal is stable. Around  $t = 9$ , the classes differ noticeably as the prominent valley is delayed for both individuals. This difference is difficult to see without using the reference NF class.

Table 2. Classes

| Label | Letter | Description  |
|-------|--------|--|
| 1     | NF     | No-fault   |
| 2     | T      | Thicker drill steel.   |
| 3     | A      | A-seal missing. Leakage from high pressure channel to control channel. |
| 4     | B      | B-seal missing. Leakage from control channel to return channel.        |
| 5     | R      | Return accumulator, damaged.   |
| 6     | S      | Longer drill steel.  |
| 7     | D      | Damper orifice is larger than usual.                                   |
| 8     | Q      | Low flow to the damper circuit.  |
| 9     | V      | Valve damage. A small wear-flat on one of the valve lands.             |
| 10    | O      | Orifice on control line outlet larger than usual.                      |
| 11    | C      | Charge level in high pressure accumulator is low.                      |

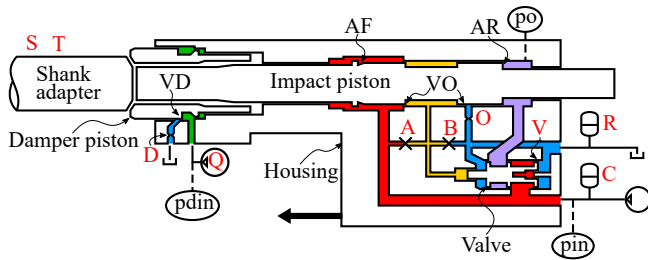


Figure 3. Schematic image of the percussion and damper system including the different induced fault modes in red capital letters, and sensor locations in ovals.

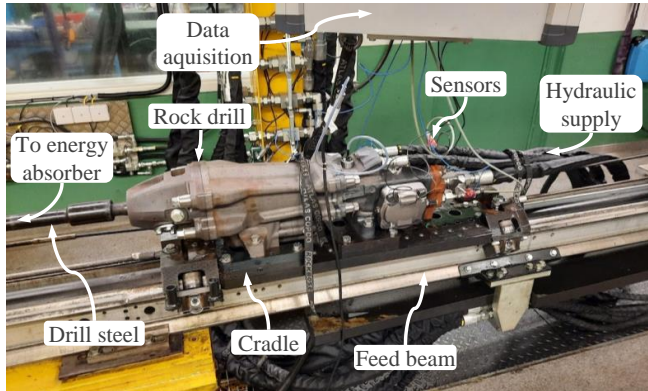


Figure 4. The test setup used to collect data.

#### 4. PRE-PROCESSING

A small amount of processing is required to adapt the data for the classification task. The data is normalized and divided into separate impact cycles.

##### 4.1. Normalization

Data is normalized for each cycle according to:

$$x_{norm} = \frac{x - \mu}{\sigma} \quad (1)$$

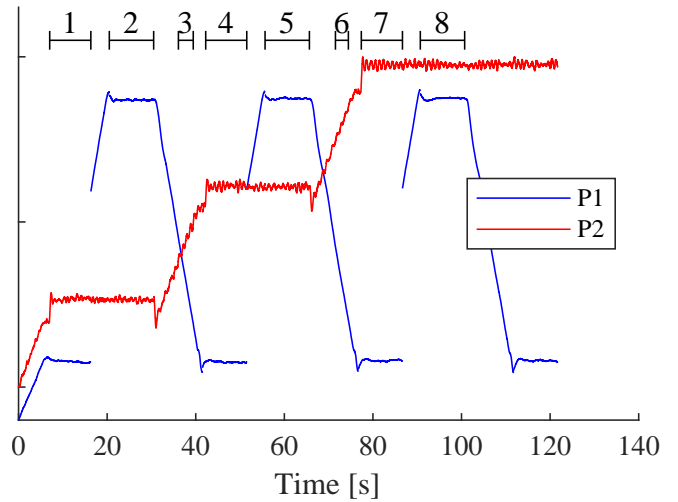


Figure 5. Two variables P1,P2 are changed to emulate different individual setups. Black markers indicate where different individuals 1-8 are collected.

where  $x$  is a pressure time series from a single cycle,  $\mu$  and  $\sigma$  is the mean and standard deviation of  $x$  respectively. This is done for two reasons. First, proprietary reasons to anonymize the exact levels of pressures within this particular machine. Second, the mean levels and standard deviation are not the main carrier of information regarding the different faults, but are rather a result of the control systems attempts to regulate the pressures and flows.

##### 4.2. Division into cycles

The collected, several seconds long time series are divided into individual impact cycles. In this way, a large number of samples are generated without the need to run hundreds of experiments in the test cell. This split is done at the time of impact to capture the main stochastic element of the data generation, the response from hitting the shank adapter and possible reflections from the rock affecting the return velocity of the piston.

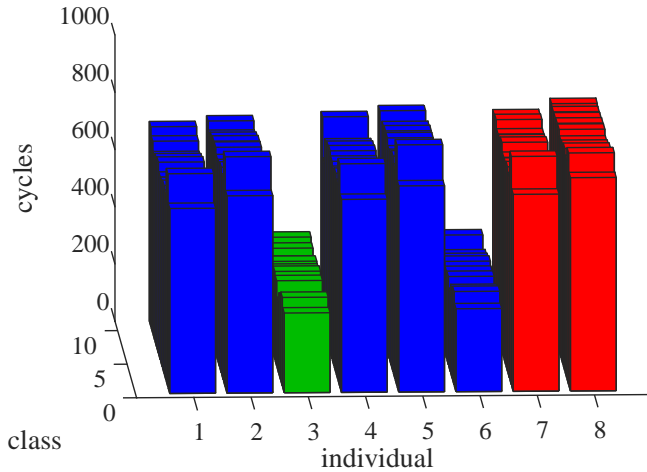


Figure 6. Number of cycles for each class and individual. Blue represents training data, green is data used for online scoring and red a holdout test set used for scoring.

## 5. STRUCTURE OF DATA FILES

Data is supplied in comma separated text files for maximum portability. Time series data from a certain sensor, for a particular individual is stored together with labels in separate files according to the naming convention: *data\_{signal}\_{individual number}.csv*. Each individual is also accompanied by a number of reference samples collected from the No-fault class from the same individual. Such file is named according to *reference\_{signal}\_{individual number}.csv*. Each row in the .csv file contains one impact cycle. The first value in each row is the *Label* according to Table 2.

The order of the different cycles are identical between the different sensors. Hence, the first row in file *data\_po1.csv* is collected from the exact same cycle as the first row in *data\_pin1.csv* etc. Combining the three files from the same individual enables multivariate time series classification. The length of the data section varies between different cycles.

### 5.1. Training/validation split

As a result of how parameters P1 and P2 are varied to emulate the individuals, some relations between the individuals exist. Mainly, the varied parameter P2 has a larger influence on the pressure signature than parameter P1. It is therefore easier to generalize between certain classes. If there is data from all available individuals in the training split, the classification problem is simplified compared to the test scenario. To avoid this, the individuals used in training/validation should be different. One example would be using cross validation where one split could be training on parts of individuals 1, 2, 6, then evaluating on 4, 5 etc.

## 6. CHALLENGE FORMULATION

This section describes the rules as used during the PHM data challenge 2022. Given are multivariate time series of sensor readings, and their corresponding class from a fleet of eight individuals. Five individuals are supplied as training/evaluation including labels. One individual is used for online scoring with a limited number of evaluations per day. Supplied are also reference measurements from each training/evaluation individual.

### 6.1. Evaluation metric

Evaluation of the model performance will be carried out using an **independent test set** that is being released for one-time assessment at the end of the data challenge. The reference measurements belonging to the test set individuals will also be made available at this time. The metric of evaluation is:

$$\text{Accuracy} = \frac{\text{Correctly classified cycles}}{\text{Total number of cycles}} \quad (2)$$

## 7. CONCLUSION

The collection and properties of a time series fault classification data set are presented. This data is made public for the benefit of the research community and industry, and is hoped to generate both interesting new classification techniques, and trigger new thoughts on how reference data from specific individuals can be utilized.

## ACKNOWLEDGMENT

The authors would like to thank Peder Haraldsson and Martin Persson, Epiroc Rock Drills AB, for assistance during data collection in this work. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg foundation.

## BIOGRAPHIES



**Erik Jakobsson** was born in Mjölby, Sweden in 1987. He received a PhD degree in 2022 from Linköping University, Linköping, Sweden. He is currently employed at Epiroc Rock Drills AB as a condition monitoring specialist. His main research interests are related to condition monitoring, sensors in harsh mining environments and the logging and use of on-board sensor data to improve the use and product development of mining machinery.



**Erik Frisk** was born in Stockholm, Sweden, in 1971. He is currently a Professor with the Department of Electrical Engi-

neering, Linköping University, Linköping, Sweden. His main research interests are model and data-driven fault diagnostics and prognostics optimization techniques for autonomous vehicles in complex traffic scenarios and.



**Robert Pettersson** was born in Askersund, Sweden in 1981. He received his M.Sc. in Vehicle Engineering at Royal Institute of Technology, Stockholm, Sweden, in 2006, and the Ph.D. degree in 2012 in Engineering mechanics, Mechanics department, at the same institute. Since 2012 he is employed at Atlas Copco Rock Drills as Specialist in Applied Mechanics. The research interest from the time as Ph.D. student is mechanics and optimization with application to multibody systems. As a professional within mining machinery development, his main interest is in general

mechanical analysis and methods for basic development and in his current position he works with e.g. structural analysis (Finite Element Modelling), load measurements and evaluation, analysis of Mechanical Rock Excavation specific topics, stability, and reliability methods.



**Mattias Krylander** was born in Linköping, Sweden in 1977 and is an associate professor at the Department of Electrical Engineering, Linköping University, Sweden. His research interests include model based diagnosis and prognosis and autonomous vehicles. As a way to cope with the complexity and size of industrial systems he has used structural representations of models and developed graph theoretical methods for assisting design of diagnosis systems and for fault isolation and sensor placement analysis. In recent years he has developed data driven techniques for diagnosis and prognosis applications.

structural representations of models and developed graph theoretical methods for assisting design of diagnosis systems and for fault isolation and sensor placement analysis. In recent years he has developed data driven techniques for diagnosis and prognosis applications.