# Anomaly Detection and Diagnostics in Distribution Networks Using High-Frequency PQ Data

**Xu Jiang**

A thesis submitted for the degree of Doctor of Philosophy to Department of Electronic and Electrical Engineering

University of Strathclyde

March 2020

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

# Abstract

Fault analysis based on high-resolution data acquisition is growing in use as it offers a more complete picture of faults, which provides an opportunity to deal with failures more effectively. However, with increasing volumes of data being collected, it becomes impossible for engineers to interpret every fault instance. To solve this, this thesis proposes novel power network fault detection and diagnosis methods applied to continuous high-frequency Power Quality (PQ) data. These novel methods deliver online anomaly segmentation, fault classification, and automatic fault labelling. The work addresses the need for increasing levels of situational awareness in distribution networks and its corresponding data-related challenges. The combination of these contributions can achieve automatic extraction of information from operational PQ data without excessive manual effort. This research uses simulated cases and operational data to validate the effectiveness of the contributions. The significance of this research is that it extracts critical information from continuous PQ data streams and automatically interprets the segmented signals, which reduces the demand for expert interpretation. In addition, it can operate through intensive monitoring at a single point on the network, which enhances the observability of the distribution network without installing excessive amounts of sensors.

# Acknowledgements

I would like to express my deepest appreciation to my supervisors, Dr. Bruce Stephen and Prof Stephen McArthur, for their continuous support throughout the research and writing of this thesis. Their patience, enthusiasm, encouragement and immense knowledge was contagious and motivational for me.

I would also like to thank Prof Brian Stewart from whom I have learnt a great deal. His kindness, constructive suggestions, encouragement, and optimism have been of great value to me during the research as well as for future life. And I would thank Dr Graeme West who introduced me into this great research group.

I would like to thank my colleagues in the University of Strathclyde, who have been really kind and helpful all this time. Besides, I would like to thank all my friends for the time spent with me, for the sleepless nights we were working together, and for all the fun we have had in the last few years.

Last but not the least, I would like to thank my parents and my girlfriend for their unconditional and faithful love, support and encouragement. This thesis is dedicated to them.

# Table of Contents

# Figures

# Tables

# Glossary of Terms

| | |
|---|---|
| AC | Alternating Current |
| ACC | Overall Accuracy |
| ACF | Animal Contact Faults |
| AD | Anderson-Darling test |
| ANN | Artificial Neural Network |
| BCP | Bayesian Change Point |
| CBM | Condition Based Maintenance |
| CI | Customer Interruption |
| CML | Customer Minutes Lost |
| CM | Corrective Maintenance |
| DBN | Deep Belief Network |
| DFA | Distribution Fault Anticipation |
| DFR | Digital Fault Recorder |
| DNO | Distribution Network Operator |
| DOE | Department of Energy |
| DPR | Digital Protective Relay |
| DTW | Dynamic Time Warping |
| ECF | Equipment Caused Faults |

| | |
|---|---|
| EMC | Electromagnetic Compatibility |
| EMI | Electromagnetic Interference |
| ERPI | Electric Power Research Institute |
| FICS | Fault Induced Conductor Slap |
| FN | False Negative |
| FP | False Positive |
| HIF | High Impedance Faults |
| IED | Intelligent Electronic Device |
| IEEE | Institute of Electrical and Electronics Engineers |
| JB | Jarque-Bera test |
| KF | Kalman Filter |
| KW | Kolmogorov-Smirnov test |
| KLD | Kullback-Leibler Divergence |
| L | Lilliefors test |
| LCT | Low Carbon Technologies |
| LDA | Latent Dirichlet Allocation |
| LSF | Lightning Strike Faults |
| MAVSA | Mean Absolute Variation in Squared Amplitude |
| MSEC | Mid-South Synergy Electric Cooperative |

| | |
|---|---|
| NAB | Numenta Anomaly Benchmark |
| NN | Nearest Neighbour |
| Ofgem | Office of Gas And Electricity Markets |
| PC | Personal Computer |
| PM | Preventive Maintenance |
| PMAR | Pole Mounted Auto-Recloser |
| PQ | Power Quality |
| PV | Photovoltaic |
| QQplot | Quantile-Quantile plot |
| RMS | Root Mean Square |
| ROC | Receiver Operating Characteristic |
| SCADA | Supervisory Control And Data Acquisition |
| SNR | Signal-to-Noise Ratio |
| SW | Shapiro-Wilk test |
| TCF | Tree Contact Faults |
| THD | Total Harmonic Distortion |
| TN | True Negative |
| TP | True Positive |
| UK | United Kingdom |

| | |
|---|---|
| US | United States |
| UKPN | UK Power Network |
| VIF | Vehicle Impact Faults |

# Chapter 1

# Introduction

## 1.1 Overview

With the growth in penetration of renewable and distributed generation, maintaining a safe and reliable power delivery in a distribution network has become a new challenge. In order to increase robustness of power distribution networks, utilities have moved to forestall operational problems by enhancing the visibility of their network with high frequency Power Quality (PQ) monitoring [1] [2]. This is motivated by the premise that faults have early stage signatures of their incipience and these manifest themselves in waveform artefacts [1]. There have been two key areas of focus: the first is to identify the anomalous behaviour of specific pieces of equipment and assets using PQ monitoring data [3] [4] [5]; the second, and less common, is analysis of network monitoring data to segment abnormal behaviour from normal signals [1] [6] [7]. Both topics are key problems within smart grids, and both would benefit significantly from enhanced online signal segmentation, anomaly detection and diagnosis - the improvements can enhance situational awareness within distribution networks, and the diagnosed results can be used to expedite maintenance and mitigate the effects of faults.

Compared to other monitoring sources, high-resolution PQ data is characterized by high-speed and high-fidelity signals which can provide higher quality records about the health of the network, which can be more informative to enhance the level of situational awareness in distribution networks [1]. However, analysis of such continuous, streaming data, is a significant challenge. Often segmentation of anomalous regions is used, but this also is challenging. [1] and [2] demonstrated an approach which uses high-resolution waveform data to find the early signatures of failures and analyse the resulting waveforms to provide more fault information for remedial decision support. The system utilised a knowledge-based technique with expert knowledge to analyse faults. However, developing knowledge-based systems to cover every eventuality can be time-consuming [2] – it takes a lot of time to manually investigate the characteristics and thresholds of each fault. A

potential solution is to utilise a fully automatic fault diagnostic system for operationally critical faults; to achieve this, a classifier with automatic feature selection is required. However, many conventional fault diagnostic methodologies utilise supervised classifiers which require a large number of exemplars for training. In addition to this, labelled fault exemplars are a rare resource for utilities, and the fault cause is, by its nature, of low prevalence; therefore, it can be challenging for utilities to prepare sufficient exemplars for every eventuality to train automatic classifiers. To solve this, an appropriate classifier is required to deal with a minimal set of exemplar situations, or the classifier can work with another classifier which can automatically label historical events.

This thesis therefore proposes a fully automatic fault analysis method which works with a high-frequency PQ monitoring system to tackle the complex problems of segmentation of the high-resolution signals, fault diagnosis with a minimal set of exemplar faults and automatic exemplar fault generation. High-frequency PQ in the thesis refers to the signals are continuously recorded at the serval kilo hertz for sampling frequency, such as 0.96kHz and 3.84kHz [8]. This overcomes some of the key barriers and problems outlined by recent works [1][6] and allows utilities to obtain maximally useful information with minimal manual effort. This work is based on the assumption of ideal hardware, architecture or communication infrastructure which has already been investigated in existing works - Distribution Fault Anticipation (DFA) [1]. This thesis takes a further step to investigate how intelligent processing of this data can lead to operational insight. In support of this, the work presented in this thesis additionally utilises simulated operational extremes and archived operational fault data to validate the effectiveness of the contributions in a practical context.

## 1.2 Principal research contributions

The main contribution of the thesis is summarised as follows:

- A novel integrated fault diagnostic system designed around high-frequency PQ data, which is assembled by an online anomaly segmentation model, an online fault diagnostic model and an offline exemplars generation model. The integrated fault diagnostic system can achieve an appropriate accuracy with a minimal set of exemplar faults. Additionally, the exemplar faults can be automatically generated by an offline fault labelling model, which can further reduce the demand for labelled faults in the model training phase.
- A novel anomaly segmentation method is proposed and developed, which can be used to pick up abnormal events from a high-frequency PQ data streaming. This method is robust to the noise of non-linear loads. Compared to the conventional methodologies, the detection accuracy is high and the false alarm rate is low. An innovative metric has been demonstrated for evaluating the quality of anomaly segmentation using streaming PQ data.
- A new approach is used to diagnose fault causes using waveform data recorded at distribution level substations. This method is built on a novel waveform similarity and associated inputs, which do not require waveform feature selection, and it can outperform conventional classifiers with a minimal set of exemplar faults.
- An innovative fault labelling method is proposed for distribution level fault records. This provides a possible way to form labelled historical fault records, which can be used to train intelligent classifiers and also increase fault diagnostic accuracy.

## 1.3 Thesis Outline

This thesis focuses on increasing situational awareness within a distribution network. It outlines the drivers and existing methodologies firstly, then introduces the proposed method including fault signal segmentation, fault diagnosis and automatic fault labelling. The structure of the thesis is outlined below:

**Chapter 2** reviews the levels of observability expected on distribution networks. In terms of the distribution network maintenance and operation, higher situational awareness can benefit power network management. The state-of-the-art methodologies are limited by various factors which will be discussed in this chapter.

**Chapter 3** develops a novel fault analysis method based on high-frequency data sampling, including the motivation, the inputs and the design. Furthermore, the various data for training and testing the proposed approach are given.

**Chapter 4** presents a high-resolution anomaly segmentation method for a distribution network, including various fault models for benchmarking. The model parameters initialization and updating are discussed, and the associated case studies with simulated extremes and archived operational data are discussed.

**Chapter 5** proposes a novel similarity-based classifier to identify fault causes trained with minimal documented faults. The performance is compared against some popular classifiers in recent literature and its practical implementation is discussed.

**Chapter 6** demonstrates an innovative approach to automated labelling of historical faults by using associated maintenance records. This uses a topic/word distribution model and is of practical benefit as it has the potential to generate training exemplars for supervised classifiers. The workflow and the implementation are discussed.

**Chapter 7** unifies the contributions into an end-to-end demonstration of the integrated fault analysis method on operational data and proposes corresponding future improvements.

**Chapter 8** summarises the contributions and the implementation of the research. Furthermore, future work, such as developing associated measurement devices and a data processing platform, are discussed.

## 1.4 Publications

**Journal papers:**

- Extracting distribution network fault semantic labels from free text incident tickets B. Stephen, X. Jiang, S. D. J. McArthur, *IEEE Transactions on Power Delivery*, PP(99):1-1, October 2019, Available as IEEE Xplore early access; DOI: 10.1109/TPWRD.2019.2947784

- Automated Distribution Network Fault Cause Identification with Advanced Similarity Metrics, X. Jiang, B. Stephen, S. D. J. McArthur, *IEEE Transaction on Power Delivery*, PP(99):1-1, May 2020, Available as IEEE Xplore early access; DOI: 10.1109/TPWRD.2020.2993144

- A Sequential Bayesian Approach to Online Power Quality Anomaly Segmentation X. Jiang, B. Stephen, S. D. J. McArthur, *IEEE Transaction on Industrial Informatics*, PP(99):1-1, June 2020, Available as IEEE Xplore early access; DOI: 10.1109/TII.2020.3003979

**Conference papers:**

- Automated fault analysis and diagnosis using high-frequency and maintenance data from distribution networks
  X. Jiang, B. Stephen, S. D. J. McArthur, The 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), Bucharest, Romania, 2019

# Chapter 2

# The State-of-The-Art for Observability in Distribution Networks

## 2.1 Introduction

Power networks can be categorized as three parts: generation, transmission and distribution level. They can be used to transmit electricity from large generation plants to distribution network loads. Distribution networks have higher customer interaction than transmission networks, which means that the electrical components in the distribution network can have more uncertainties in their operating states that pose a greater range of potential fault causes [9], such as tree contact, animal contact and equipment degradation. Furthermore, with the power system trend to be more decentralized [10] [11], and more new low-carbon technologies being embedded, it becomes more important to enhance the situational awareness in distribution networks, which can help DNOs reduce the penalties associated with excessive CI and CML by acting on faults before they evolve into outages [1][12][13]. Therefore, distribution networks need more advanced techniques to achieve this. This chapter outlines the motivation of increasing situational awareness for distribution networks, and two corresponding approaches are investigated. Section 2.2 introduces the benefits of increasing situational awareness, including reducing financial penalties and enhancing maintenance efficiency. Section 2.3 compares the advantages and disadvantages of existing approaches and it also demonstrates some case studies of using these methodologies.

## 2.2 The Motivation behind Increasing Situational Awareness in Distribution Networks

### *2.2.1 Visibility of the Present Distribution Networks*

As Fig 2-1 shows, a distribution network usually connects to a transmission network through Grid Supply Points (GSPs) [14], then distributes electricity to the customer side via overhead lines or underground cables. Furthermore, a distribution network usually has many assets, such as transformers, circuit breakers, sectionalisers and lightning arrestors, etc., dispersed across the network. Any fault happening on the assets can affect the

network's operation and reliability. While existing tools such as State Estimators [15] exist for understanding network parameters under normal operating conditions, it is also important to identify the network operating parameters reflecting asset condition under adverse conditions. Traditionally, the control room identifies abnormalities in distribution networks through observing breaker movements at higher voltage levels significant failures or receiving calls from customers who are off supply; they will then send crews to investigate the fault cause and take remedial action. Generally, conventional fault detection is passive and fault causes are identified through manual analysis of weather and fault behaviour [16].

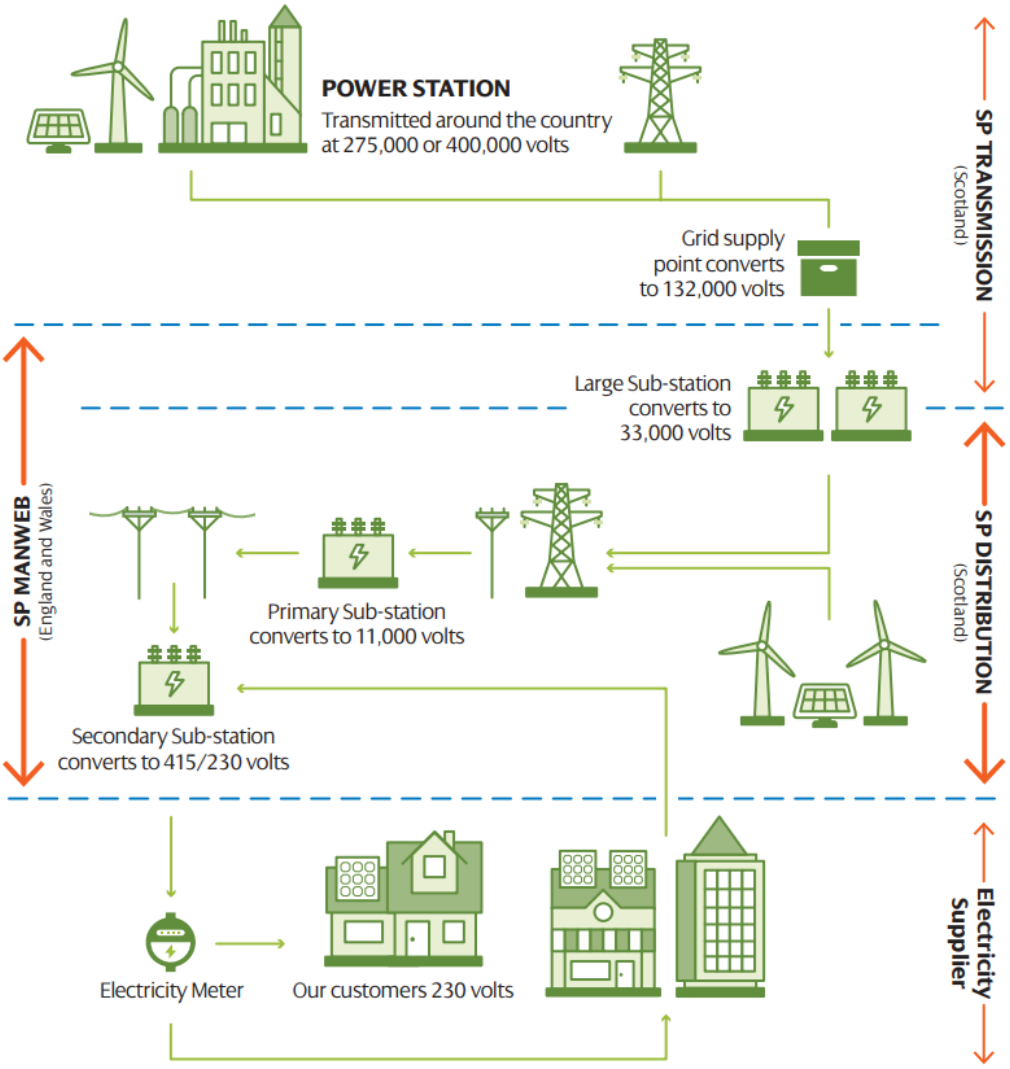



Fig 2-1. Example of Power System hierarchical structure [17]

### *2.2.2 Financial Penalties and Reliability Indices*

Many stakeholders are involved in power network management. In the UK, the Office of Gas And Electricity Markets (Ofgem), a government office for regulating electricity markets, has published a series of quality of service standards for distribution utilities to deliver [18]. Utilities can be penalized or rewarded according to their annual performance of operation and maintenance [19][20]. The commonly accepted indices are Customer Interruption (CI) and Customer Minutes Lost (CML) [18] [20] in the UK. CI indicates the number of customer interruptions per 100 customers in one year [20]. CML is the average length of time customers are without power per interruption for one year [20]. The higher levels of situational awareness of distribution networks can help DNOs reduce the penalties associated with excessive CI and CML by acting on faults before their seriousness elevates to a level where remedial work results in lengthy outages.

### *2.2.3 Maintenance Implementation of Distribution Network*

Maintenance can be classified into three categories: Corrective Maintenance (CM), Preventive Maintenance (PM) and Condition Based Maintenance (CBM). For CM, crews only take maintenance actions when some significant damage or broken signatures are found on an asset. PM works on the time-based planning which requires crews to periodically check an asset to prevent an incipient fault developing into a failure. CBM has the same goal as PM, which is to prevent an incipient fault, but it assesses the current working condition and performs maintenance only when the need arises. The British Standards document defines CBM as 'the maintenance policy carried out in response to a significant deterioration in a machine as indicated by a change in a monitored parameter of the machine condition' [21][33].

Presently, CM is treated as the first choice because its implementation is simple, and the capital cost, such as sensor installation, is low. However, since CM is a passive scheme, the consequence of failure is the largest among these schemes. Therefore, CM is usually

conducted for the non-critical assets in a distribution network. PM is usually conducted taking account of a historical degradation model. For example, PM is usually conducted in spring or autumn because they normally have higher risks from heavier rain than other seasons. However, PM only depends on the historical degradation which does not take into account the present condition. Therefore, this approach is not be the most cost-efficient approach. CBM is a methodology taking account the present condition, which can prevent a fault before evolving into a failure. Although CBM can limit the consequence of failure to the lowest among three schemes, its capital cost is indeed the highest. Therefore, CBM is now only employed into some critical distribution assets in distribution level, such as transformers [22] and circuit breakers [23]. Generally, the maintenance of distribution networks is becoming more proactive, which requires the support of a higher level of situational awareness.

From the perspectives of finance and maintenance, distribution networks are looking forward to higher levels of situational awareness. One solution is to use an automatic fault analysis system. The next section will introduce two existing methodologies and then will compare their advantages and disadvantages.

## 2.3 Existing Fault Analysis Methodologies

This section will introduce two methodologies to increase situational awareness on distribution networks. Based on different frequencies of data acquisition, the methodologies can be categorized as conventional fault analysis and continuous high-resolution monitoring-based fault analysis. Conventional fault analysis usually utilises existing monitoring, such as Intelligent Electronic Devices (IEDs) and Supervisory Control And Data Acquisition (SCADA) systems; these devices are usually centralised to obtain data at regular intervals rather than on a continuous basis [24][25]. High-resolution monitoring-based fault analysis will usually collect continuous data at kHz sampling rates and analyse it in an online manner, which would be operate without interruption.

### *2.3.1 Fault Analysis using Existing Data*

#### *2.3.1.1 Fault Diagnosis with SCADA alarm data*

Supervisory Control And Data Acquisition (SCADA) systems have been widely deployed for data acquisition, processing and communication, which produces a large amount of low-resolution data. SCADA is usually integrated into Distribution Management System (DMS). The current role of DMS is more akin to a dashboard which only presents the data rather than analyses it. In future, DMS are expected to process more events, such as voltage violation, and work with automatic equipment, such as reclosers. SCADA provides an opportunity to automatically extract critical fault information to increase situational awareness on distribution networks. In terms of the data resources, the research concerning the use of SCADA data in fault diagnostics can be categorized into two areas: the first uses SCADA alone; the other approach is to combine SCADA with other data sources, such IED data, to enhance the analysis result.

Early research on SCADA data analysis used knowledge-based fault diagnostic approaches [26] [27] [28]. This then extended to the use of Digital Fault Recorder (DFR) data. The DFR monitored and recorded detailed disturbance data. It could provide more information on transient waveforms compared with SCADA. Therefore, fault analysis approaches for DFR data were developed [29][30]. The work advanced to using both data sources to improve the diagnosis [31][32][33][34]. Additional data resources, such as recloser data, were investigated to help analyse faults in combination with SCADA data [24].

#### *2.3.1.2 Fault analysis with IED data*

Intelligent Electronic Devices (IEDs) are used for power system automation. IED data is a common data source for fault analytics in distribution networks [35]. As the previous subsection shows, DFR is one of the IEDs utilised for fault diagnosis. Furthermore, some

other IEDs, such as Digital Protective Relay (DPR) and Pole Mounted Auto-Recloser (PMAR), were developed for fault analysis [24][31][32]. This is because DPR and PMAR can provide more fault details compared to SCADA or DFR. Generally, compared with SCADA, IED data benefits from more fault information, but it is only for a local area.

The early work on SCADA, IED, DFR and other data analysis for fault diagnosis indicated the possibility of improved situational awareness. However, these data sources could not always capture faults that were transient or intermittent in nature. Consequently, there was a move to continuous data measurement to compensate for this, which requires research to provide fault diagnosis for such data. Thus, research starts to move into continuous high-resolution fault diagnosis.

### *2.3.2 Continuous High-Resolution Fault Diagnosis*

Over last two decades, as data acquisition becomes cheaper, a new continuous high frequency fault analysis approach [1][12][13][38] is being developed to further improve situational awareness of distribution networks. Compared to the conventional fault analysis, this approach does not require the installation of many sensors for power networks, it only needs a few centralized sensors to collect data; then it can recognize the fault in the network by analysing the data. This approach can provide more fault information, especially for incipient faults, for DNOs to achieve better recognition and reaction to faults. The following subsections will demonstrate two cases of analysing continuous high-resolution fault data to increase situational awareness of distribution networks.

#### *2.3.2.1 Case Study: Cable Canary*

A large number of aged underground cables are used in London which makes the fault management and the corresponding maintenance difficult. The local DNO, which is UK Power Networks (UKPN), proposed a novel solution to this by analysing the high frequency signals acquired from the medium voltage distribution network [38]. This

project was named Cable Canary which is aimed to reflect the early warning capability of the technology. In this project, Radio Frequency Current Transformer (RFCT) from EA technology is used to obtain signals at 200kHz to 20MHz [39], which can collect 4800GB data for multiple sensors per day, which is impossible to manually process. However, some of the data is useful for preventing a fault evolving into an outage which is difficult to detect by the low-frequency fault analysis approaches. One of the examples is demonstrated in Fig 2-2.

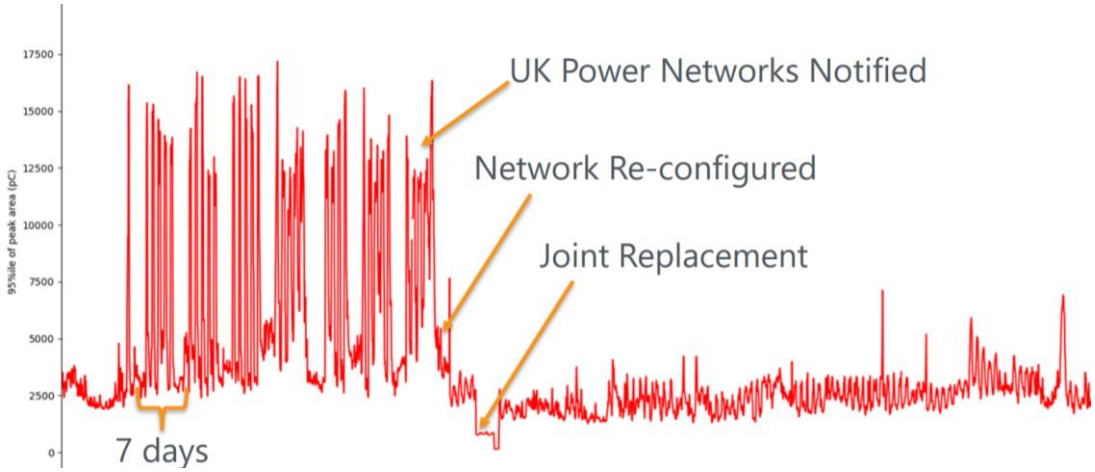



Fig 2-2. Successful case of detecting incipient fault with a high-frequency sensor in the Cable Canary project [38]

As Fig 2-2 shows, the high frequency signals found a significant high partial discharge at a very early stage. However, this abnormality was not be seen by a conventional device until more than a month later. Although continuous high frequency signals can provide more information, the volume of data can be a challenge for processing. The next step of this work is to build automatic fault analysis to extract key information from the volume of data. The next case study will demonstrate an application of using continuous high frequency signals to diagnose faults online.

### *2.3.2.2 Case Study: DFA*

Distribution Fault Anticipation (DFA) is a commercial waveform-based fault analysis product for increasing situational awareness of distribution networks [1][12][13]. This application installed current and voltage measurement instruments on key feeders of the substations. The measurement devices are shown in Fig 2-3.


Fig 2-3 The DFA measurement device for fault detection and diagnostic device [40]

The input signals include three-phase voltage, current, relevant weather data input and communication ports. The device captures the waveform data stream at a rate of 20 samples per AC 50 Hz cycle [41]. Therefore, it is implied that the fault diagnostics are built on the waveform data and the associated weather data. This work can identify a fault occurring downstream from a distributed electronic device [1]. The DFA device does not require extensive communication between devices, and this is achieved over an internet connection. Based on the hardware measurements, DFA identifies vegetation contact,

cable failures, switch and clamp failures, lightning arrestor failure, repetitive overcurrent faults, voltage regulator failure and capacitor problems [1][40].

DFA utilised a Fuzzy Logic methodology with expert knowledge to identify the faults [2]. In this way, Fuzzy Logic does not require numeric data for knowledge learning, but it does need a lot of time to manually capture knowledge for different faults, especially for low-prevalence faults; this results in a limitation of the type of fault identified. The improved methodology extracts shape based features from the RMS waveform and then uses Fuzzy Logic to calculate the probability that the features conform to each fault signal features in theory [2].

Two successful cases are discussed: Fault Induced Conductor Slap (FICS) fault and clamp failures on transformer case.

**Fault Induced Conductor Slap Detection:**

DFA was used to help one of the US utilities, Mid-South Synergy Electric Cooperative (MSEC), to identify faults [42]. MSEC instrumented ten circuits with DFA technology. One of the successful examples of detecting FICS is demonstrated in Fig 2-4 and Fig 2-5. FICS is a complex phenomenon which usually occurs when the initial fault causes the overhead line to swing and make contact with another cable, and then causes a second fault [42].

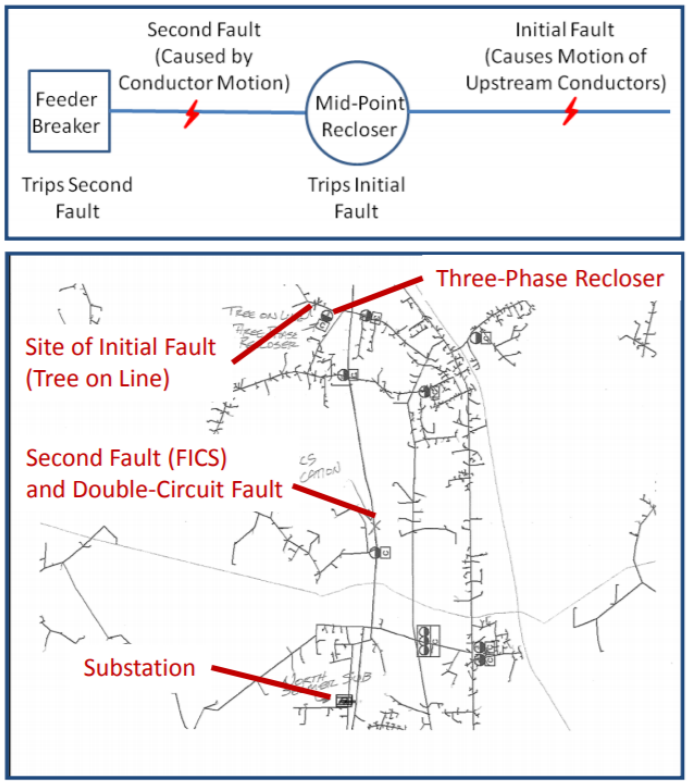



Fig 2-4 FICS fault detection by DFA [40]

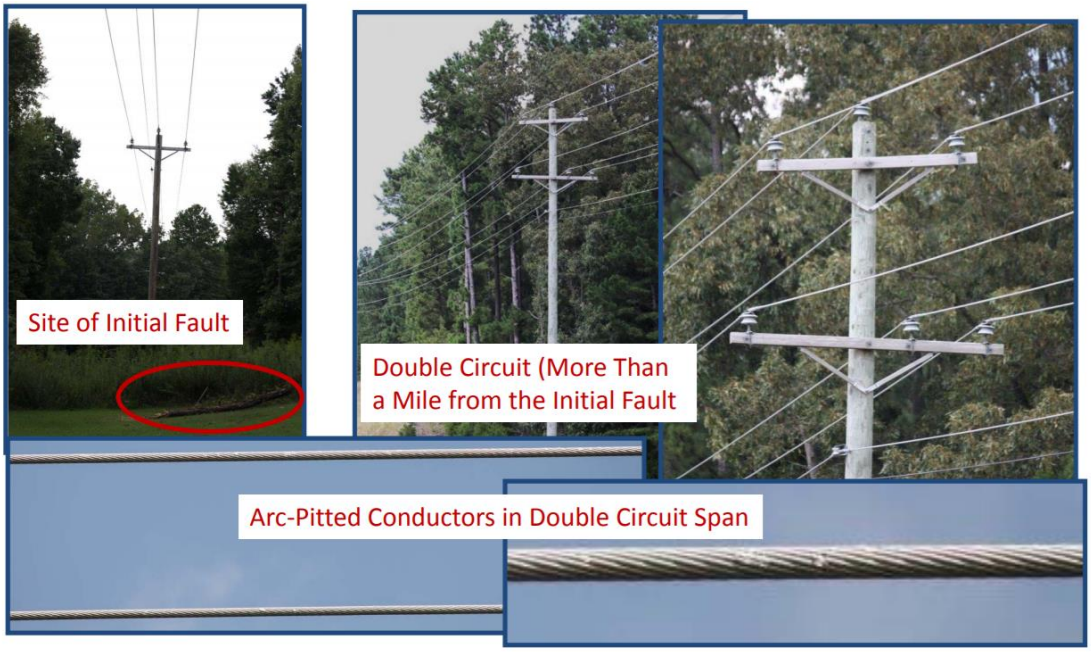



Fig 2-5 Scene of FICS [40]

As demonstrated in Fig 2-4 and Fig 2-5, a tree contacted an overhead line which results in an initial fault. After a while, the overhead line swing caused by the initial fault results in the second fault, and the second fault is more dangerous because it is closer to the substation. This fault is difficult to detect by conventional methodologies; however, DFA has identified this fault through analysing high-frequency signals and provides an alert before a failure occurs.

**Clamp failure:**

Clamps can be critical in networks. They can be used to connect cables and transformers. However, the identification of a clamp related fault is difficult for conventional methodologies. This is because a clamp failure only makes a subtle change to the waveform. An example is demonstrated in Fig 2-6.

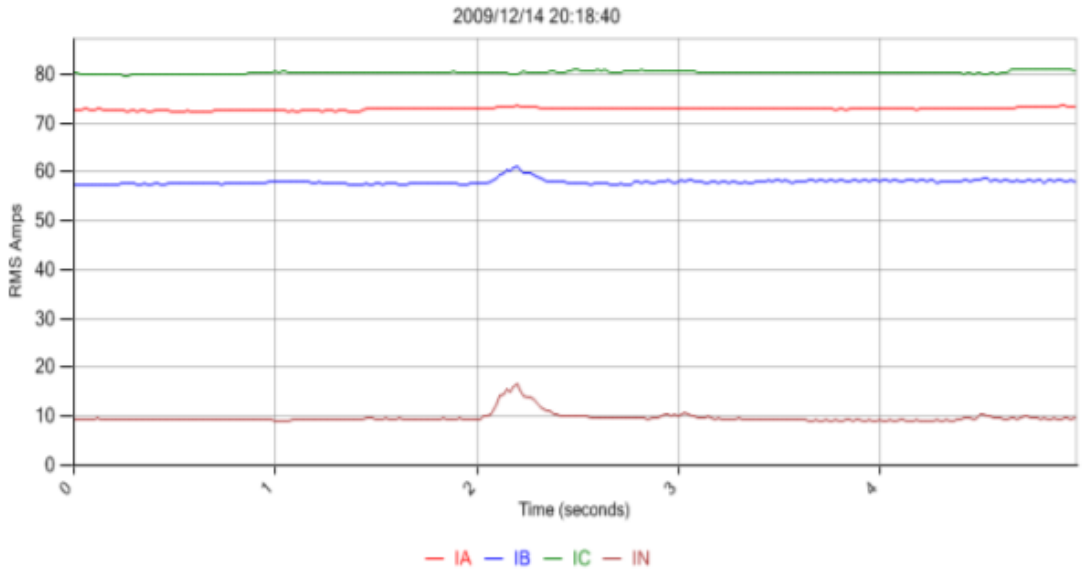



Fig 2-6 The waveform of a clamp failure in distribution networks [40]

Although the electrical variance caused by a clamp is minor, DFA detected it in a distribution network: a clamp on a transformer failed over 21 days before it was corrected. During this period, the abnormal waveform recurred for 2333 times and local customers

experienced four outages, demonstrated in Fig 2-7. However, DFA detected it and sent alerts ahead of the first outage.

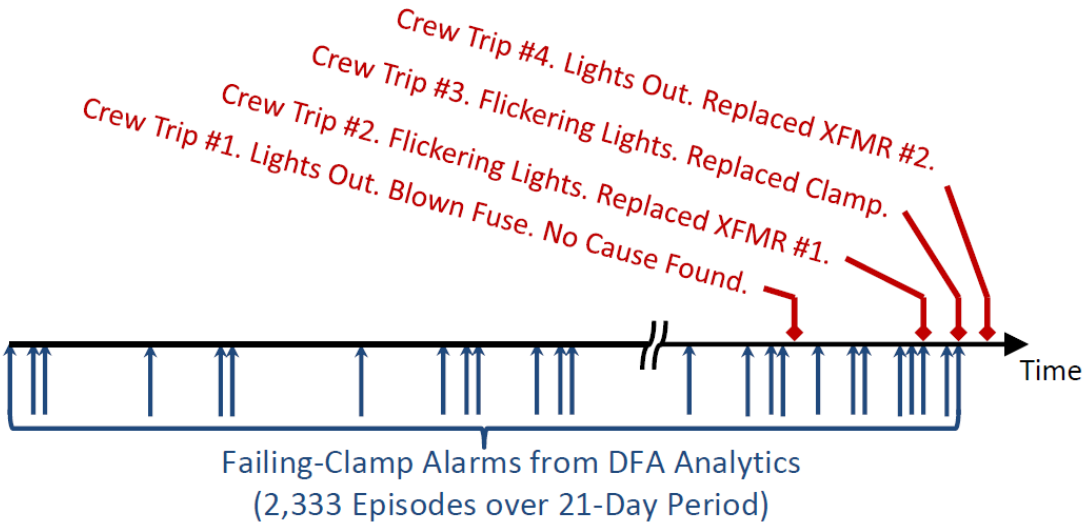



Fig 2-7 The procedure of the conventional maintenance and DFA records [40]

Generally, a continuous high-resolution fault diagnosis can provide more information for failure, which can help DNOs to expedite remedial maintenance and reduce the effect of the faults. However, the state-of-the-art is built on expert knowledge, which can be time-consuming to obtain, and which is impossible to cover every eventuality for every distribution network. To solve this, a data driven approach will be proposed in this thesis.

## 2.4 Challenges

Knowledge-driven methodologies cannot identify all faults, especially for the faults which cannot be easily characterized and quantified in terms of voltage, current and other PQ measures. Therefore, a data-driven approach is needed to increase situational awareness of distribution networks. As opposed to a knowledge-based approach, a data-driven approach needs to automatically learn expert knowledge from measured signal data without expert intervention. Combining this with the characteristics of high-frequency data, this results in a number of data related challenges as outlined below:

- Excessive volumes of input data:

As Section 2.3.2.1 shows, the continuous high-frequency data set can be extremely large [38]. In this data, most of the data are for normal operational situations, which can result in large volumes of needless storage. Therefore, to efficiently utilise storage, the continuous high-frequency data must be analysed to assess their usefulness.

- Automatic abnormal waveform interpretation:

As Section 2.3.2.2 shows, abnormal waveform needs to be automatically interpreted, this is because abnormal events are excessive, which is difficult to manually identify. Knowledge-based models need to quantify expert knowledge manually [1][12][13], while data-driven models need to learn knowledge automatically from data. Therefore, the latter model requires a large amount of labelled data for a pre-trained model; however, most DNOs would consider archiving curated fault data, marked up with diagnostic labels, to be beyond their usual remit. Therefore, it is impossible to gather such amounts of labelled data in a short time – the demand for a large number of labelled fault data sets is challenging for the data-driven approach.

Therefore, this thesis will describe the research that addressed these challenges and proposes a novel approach to solving them and delivering automated analysis.

## 2.5 Conclusion

Distribution networks are becoming more and more complex; however, situational awareness on these networks is still low, which results in DNOs being unable to make a correct decision in a timely manner. To address this, some fault diagnosis methodologies have been proposed; however, all the existing methodologies have limitations in practical implementation. In the following chapters, a new high-frequency fault diagnosis approach will be proposed, which addresses these limitations.

# Chapter 3

# Design of an Automatic High-Resolution Data Fault Analysis System

## 3.1 Introduction

Fault analysis based on Power Quality (PQ) monitoring involves continuously capturing voltage and current measurements at waveform level resolution, such as 0.96 kHz and 3.84 kHz sampling frequency, then segmenting the abnormal waveform out and diagnosing the fault based on these waveforms; therefore, the waveform of the events is critical for the fault analysis. This section will review the information carried by different waveforms which can assist fault analysis and outline the overview of the proposed fault analysis system. Section 3.2 demonstrates the motivation of automatically identifying the fault cause through analysing the prevalence of fault causes in the UK for 2015. Section 3.3 shows examples of how different faults may manifest in waveform data, including the faults caused by external effects, such as tree contact, and faults caused by internal effects, such as aging. Section 3.4 explains the associated labels for fault waveform data as well as the challenge and solution around labelling data. Section 3.5 introduces the design of the proposed automatic high-resolution fault analysis. Section 3.6 demonstrates the simulated operational extremes and archived operational data which are used for validating effectiveness of approaches.

## 3.2 Fault Cause Statistics

According to Eaton's 2015 UK power outage data [43], 640 outages were found which affected 2.56 million people. Among these outages, damaged equipment or human error takes 51.56% of all, as Fig 3-1 shows. The root cause of approximately 30% faults are unknown, and approximately 18% of faults are caused by external effects such as weather and trees. As section 2.2.2 mentioned, excessive CI and CML can result in the utilities being punished. Furthermore, from an operational perspective, knowing the broad cause of a fault prior to going into the field to investigate would inform maintenance crews of how to equip themselves and how to formulate a plan for finding the fault cause and instigating remedial action. Such fault information has the potential to shorten the

timescales in which this restoration plan may be executed. An example would be in distinguishing an overhead line bird strike from a vehicle hitting a pole – the pole impact necessitates visual inspection to confirm the cause whereas the bird strike is transient in nature and therefore pointless to look for – hence restoration of power can be immediate. To solve the problem, utilities are looking for a higher observability through a continuous high-frequency fault analysis to identify the fault causes for downstream network, then the results can be used to accelerate the recovery procedure.

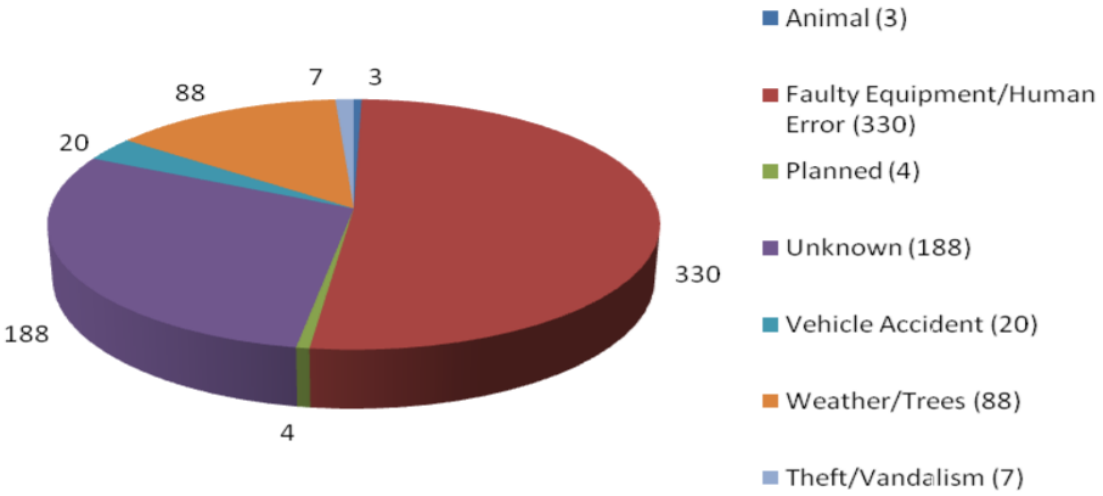



Fig 3-1 Fault causes statistic for UK networks [43]

### 3.3 PQ Waveform Characteristic of Different Fault Causes

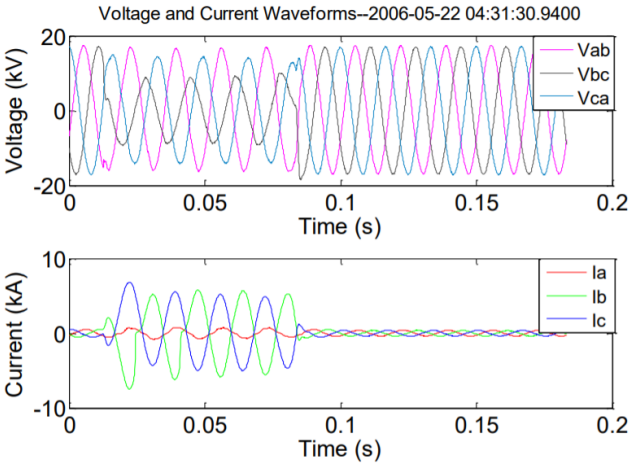
Faults or abnormal events normally can be categorized into two classes: the external faults and internal faults. External faults are usually caused by environmental incidents, such as tree contact, animal contact, vehicle impact and lightning strikes. The internal faults usually result from damage to, or defects in, the equipment. If a fault occurs, it usually introduces specific characteristics into the PQ waveform. For example, a short-circuit usually results in an abrupt overcurrent, and then the waveform distortion will affect the quality of power delivery. However, the waveform distortion provides an opportunity to identify faults by the characteristics of the waveform. The remainder of this section will

demonstrate the characteristics and associated expert knowledge for different faults through visualizing some realistic operational data.

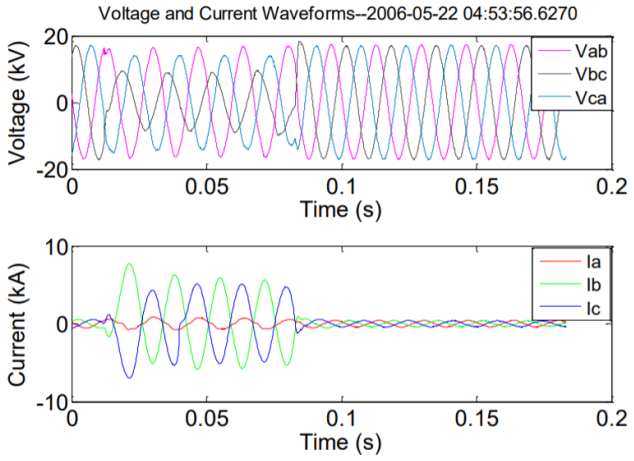
### *3.3.1 External Faults*

#### *3.3.1.1 Tree Contact Fault*

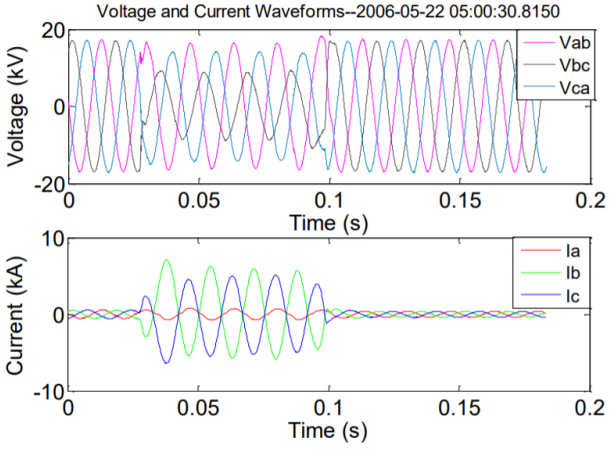
Tree contact faults are harmful and can affect the operation of power networks, some of which may even threaten the safety of the public. Furthermore, this fault is difficult to detect using conventional means. Tree contact faults are usually associated with overhead lines. These failures usually manifest as short-circuits; however, the circuit current may be not sufficiently large to trigger the protection system. Additionally, this fault usually can be produced by swinging tree branches, which can result in subsequent faults [44]. These events can be called fault episodes. Each fault episode is very short in duration, such as one to three cycles, before it clears itself [8]. Furthermore, the tree contact faults can also produce some abrupt and significant overcurrent signatures in the waveform. This can be because the voltage and current through the tree trunk usually is slightly increased with thermal changes, and some abrupt changes manifest on the voltage signal if the branch breaks or burns [44]. Finally, some of these faults are high-impedance faults which means these fault waveforms might not be detected by conventional means, and it can be difficult to use expert knowledge to identify these faults. The following is an example of a tree contact fault episode.




(a) The first episode of the tree contact sequence




(b) The second episode of the tree contact sequence




(c) The third episode of the tree contact sequence

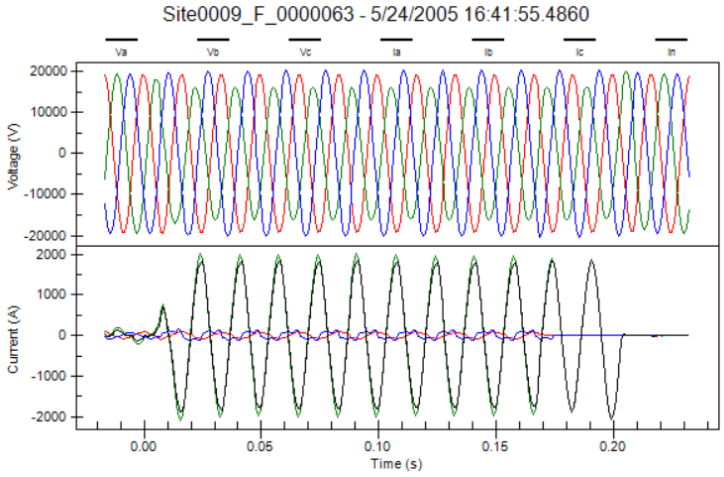
Fig 3-2 Tree Contact Fault Episodes – the same tree invoking three separate fault episodes

[8]

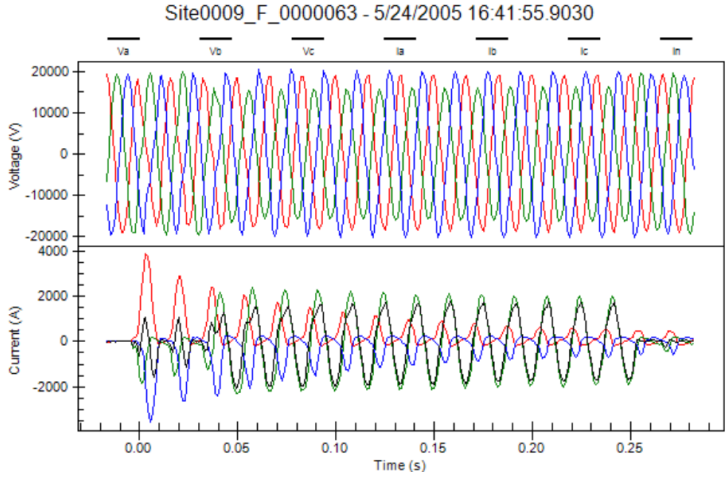
Fig 3-2 shows a sequence of faults caused by tree contact over a period of one hour. The waveform characteristics of the tree contact episodes are consistent, and each fault is cleared by a recloser. This provides an opportunity to identify the episode using shape similarity. However, each contact could damage and burn the overhead line which requires crews to find them as soon as possible. Furthermore, the episodes in Fig 3-2 are phase-to-phase overcurrent faults, which indicates the tree may affect two lines at the same time.

#### *3.3.1.2 Lightning Induced Fault*

Lightning can directly strike power network equipment, resulting in the interruption of operation or it can induce an overvoltage which damages the equipment. For physical lightning contact, strikes occur in areas with high towers usually associated with high voltage transmission [45]. Therefore, induced lightning strikes are more common on the distribution network. For induced lightning, the PQ waveform usually is coupled to strong electromagnetic fields from the lightning which can be further propagated across the power network. Lightning induced overvoltage can result in maloperation of the protection system or the associated control system in distribution networks [45]. Furthermore, lightning is usually associated with tree contact faults, because trees naturally attract lightning, and the struck trees can contact the overhead line to produce faults. Therefore, the variance of the faults makes it difficult to identify. Moreover, lightning caused faults can be a fault episode, as Fig 3-3 shows.




(a) The first episode of the lightning sequence




(b) The second episode of the lightning sequence

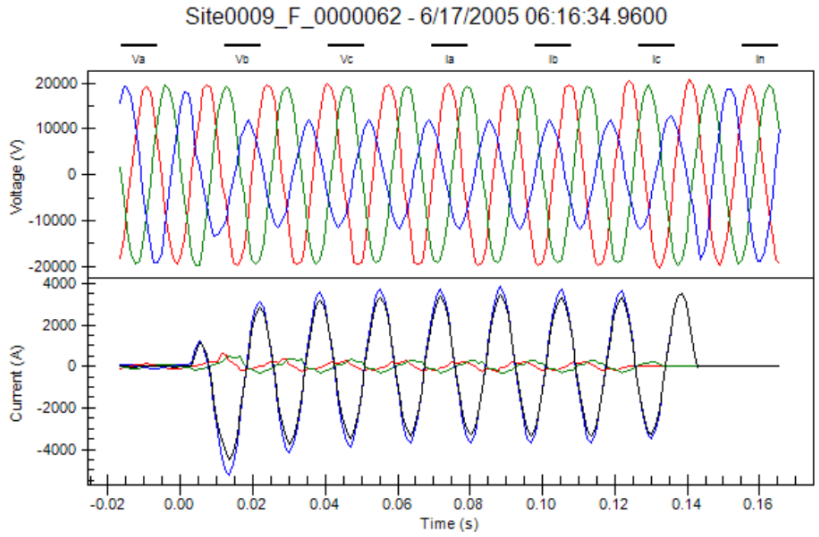
Fig 3-3 Lightning Strike Episodes [8]

As Fig 3-3 shows, the initial PQ distortion is caused by a lightning strike on the transformer which results in a single-phase of current abruptly increasing. Then the strike burned the primary feeder down, which results in operational failure of the power network. Compared with the event in Fig 3-3 (a), the event in Fig 3-3 (b) contains more harmonics and higher phase imbalance level.

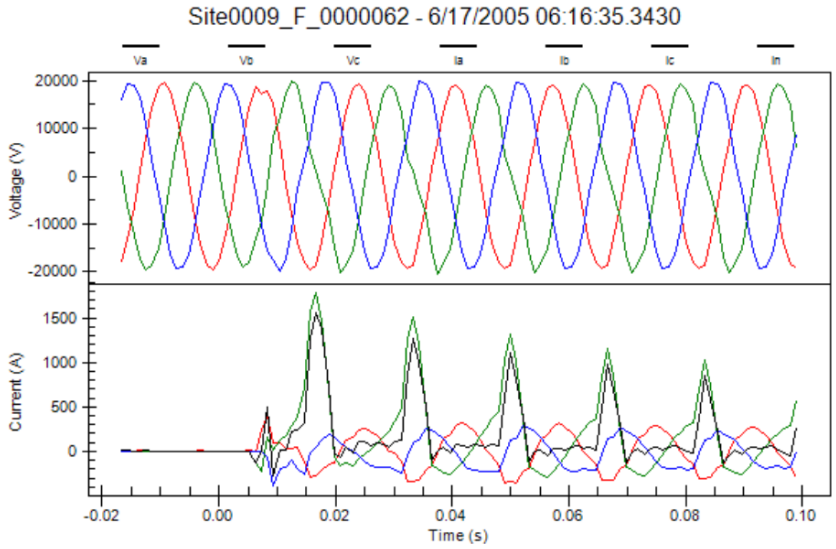
#### *3.3.1.3 Animal Contact Fault*

Animal contact faults can happen anywhere but tend to occur predominantly in rural areas. Furthermore, animal contact faults tend to happen in spring and summer. This is because animals prefer to go out in the spring and summer to hunt, rather than the winter months when they are more likely to be in hibernation. Animal faults usually manifest as

phase-to-earth or phase-to-phase short-circuit faults with current conducted through the body of the animal involved, which can result in the PQ waveform being distorted. e.g. snakes can climb on to the top of transformers, poles or towers; birds build nests on pole or towers, which can cause a short-circuit. Animals such as squirrels can also interfere with transformers [8]. These faults can result in poor power quality service which can further damage the equipment, and the fault might further threaten wildlife. An example [46] from US DoE/ EPRI Power System Events Repository is shown in Fig 3-4.




(a) The first episode of the animal related sequence



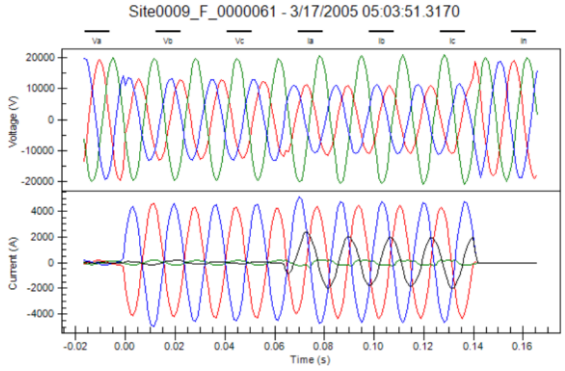

(b) The second episode of the animal related sequence

Fig 3-4 Fault episodes caused by snake contact [8]

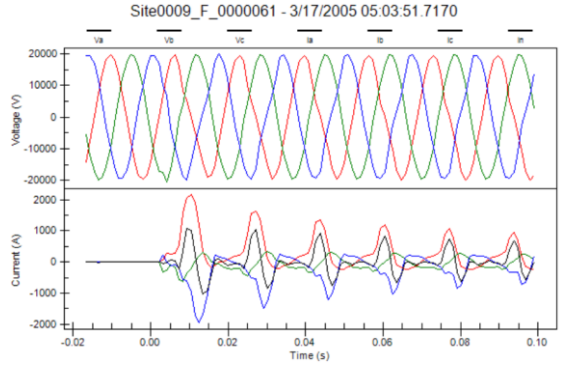
Figure 3-4 shows a fault caused by a snake touching the lightning arrester. When the snake touched the high-voltage terminal, a large current flowed through its body, resulting in a persistent ground fault and triggering the protection device. The resulting conducted fault is difficult to self-clear, therefore the duration of these faults can be relatively long or permanent.

#### *3.3.1.4 Vehicle Impact Fault*

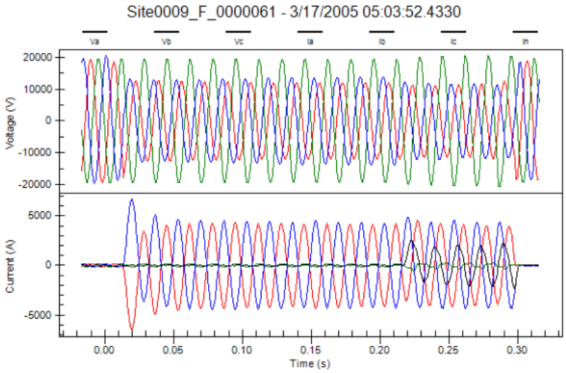
When a vehicle hits a pole, it can result in plant failure. This fault can be difficult to identify, especially in a rural area; however, it can be observed via PQ waveform distortion. The common resulting faults comprise single-phase ground faults and interphase faults [47]. Furthermore, many vehicle impact faults are permanent faults, that is they can trigger the protection devices multiple times until they are locked-out. One fault episode is demonstrated in Fig 3-5.



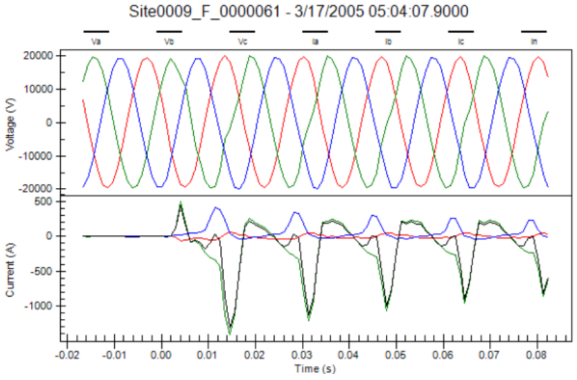

(a) The first episode of the vehicle impact sequence




(b) The second episode of the vehicle impact sequence




(c) The third episode of the vehicle impact sequence




(d) The fourth episode of the vehicle impact sequence

Fig 3-5 Fault Episodes Caused by Vehicle Impact [8]

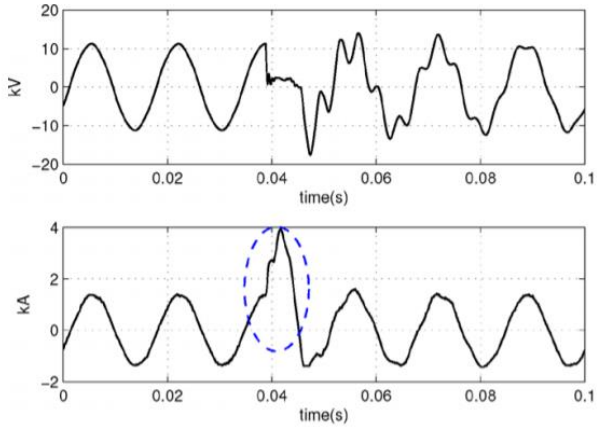
According to the weather records for that time, ice and snow were present, and the vehicle likely skidded, hit the pole, and then generated a permanent fault until the sectionalizers locked-out. Every time the sectionalizer operates, the waveform for the fault is collected. As Fig 3-5 shows, every waveform record is different which implies the fault is evolving. For example, the event from Fig 3-5(a) has more obvious harmonics and higher phase imbalance than the event 3-5(b). If the fault can be identified at the beginning, then crews can be sent to fix this fault immediately, which can prevent the fault further affecting other assets in the distribution network.

### *3.3.2 Internal Fault*

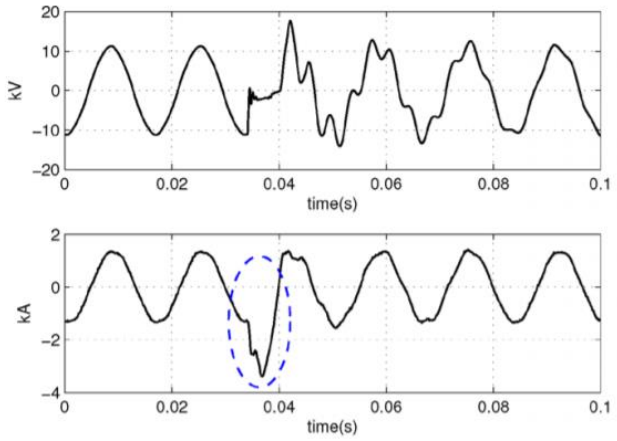
The internal fault refers to an equipment related fault, such as defects, aging and equipment damage. The following section will cover three common equipment related faults.

#### *3.3.2.1 Incipient Faults in Cables*

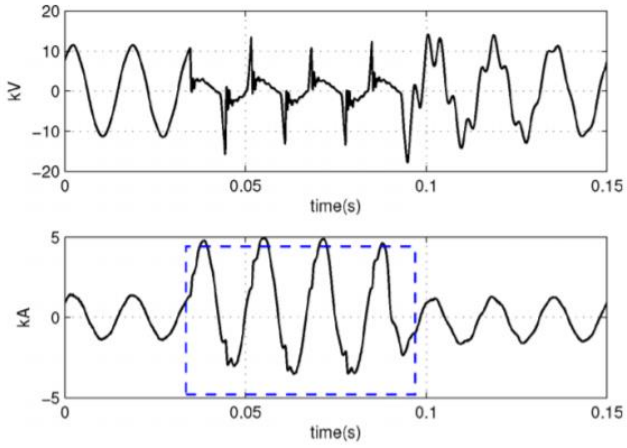
Cables have been widely used in cities or towns to transmit electricity to end users. Many underground cables in UK power networks, which have existed for over 100 years, are now at a high risk of failure. However, the observability of underground cables is low. The underground cable possesses unique waveform characteristics [3][48][4] which provide an opportunity to identify faults. The duration of the incipient cable faults can be sub-cycle or multi-cycle. Many cable faults can be self-cleared, but they can also recur for many times until they evolve into permanent faults, such as the case demonstrated in Fig 3-6.




(a) Incipient faults on 2008-11-12 at 19:40




(b) Second episode on 2008-11-12 at 21:11




(c) Permanent fault on 2008-11-14 at 15:51

Fig 3-6 an incipient fault developed into a permanent fault [3]

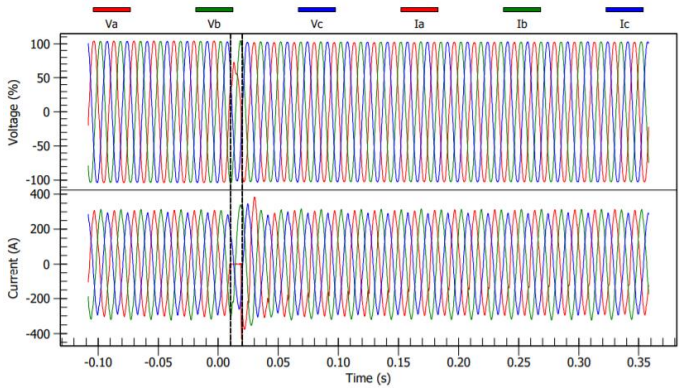
As Fig 3-6 shows, arcing caused waveform distortion of the sinusoid peak at 19:40 on the 12th of November [3]. Then the arc evolved into a permanent fault over two days. There are three waveform shots captured during the two days, and the waveform shape seems

similar to each other, especially for the incipient faults. This provides an opportunity to detect and identify faults in a timely manner, also help DNOs prevent recurring episodes.

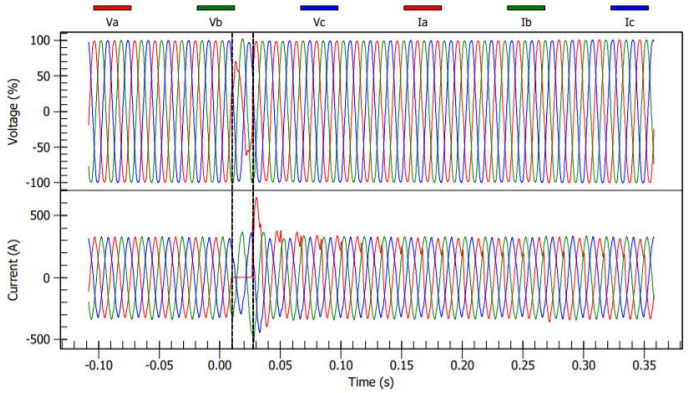
Arcing is one of the most common faults for cables. Arcing usually manifests on the waveform as transient and intermittent distortion. The arcing can usually be caused by failure of cable joints, the failure of cable terminations and defective insulation. The voltage waveform during arcing can be approximated as a square wave, as [3] shows. However, the signal captured remotely from the originating fault location can contain more noise compared with the pure fault waveform, therefore, it can be more challenging for fault identification at a distance.

#### *3.3.2.2 Incipient Fault on Transformers*

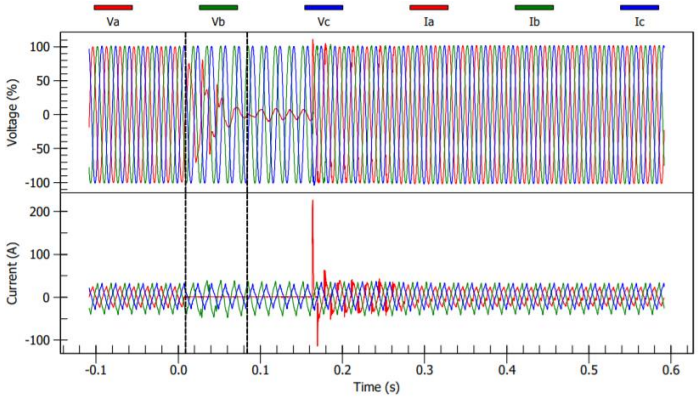
Transformers are used for increasing or decreasing the voltage level in power networks; they are usually positioned in key locations such as substations. Previous work has demonstrated that the incipient fault of transformers can recur until they develop into a failure [49][8]. One of the fault episodes is demonstrated in Fig 3-7.




(a) first episode for transformer fault sequence on April 21st, 2010 at 6:13:41.1825




(b) second episode for transformer fault sequence on April 21st, 2010 at 9:51:36.1367



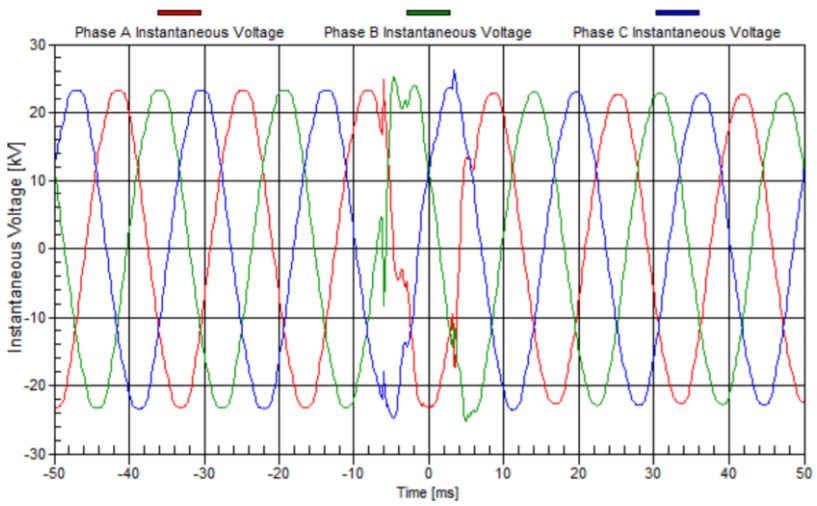

(c) third episode for transformer fault sequence on April 25st, 2010 at 8:10:43.0045

Fig 3-7 Tap Changers Failure in Transformer [8]

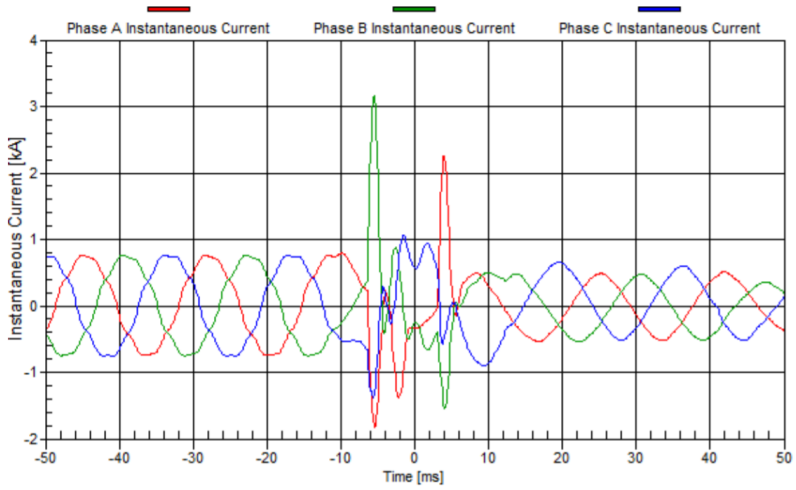
Fig 3-7 shows three waveforms during a tap changer failure. The event duration becomes longer as the fault development progresses which means the equipment is degrading. Furthermore, the current of phase A drops to zero during the disturbances. The first incipient fault is observed one week before it developed into a catastrophic failure. During this week, about 40 similar waveforms were captured [8].

#### *3.3.2.3 Capacitor Switching Restrike*

Capacitors are widely used to support reliability of networks as they can store or discharge electrical energy. Capacitor switching is a common operation for network control. Normal capacitor switching does not produce any significant transient signals; however, obvious transient signals can be observed during capacitor switching restrikes. Capacitor switching restrike has been defined as "a re-estimation of current between the contacts of a switching device during an opening operation after an interval following zero current of greater than one-quarter cycle at normal frequency" [8]. Restrikes can damage capacitors and switches, as they can cause poor power quality output. Fig 3-8 demonstrates an example of a poor power quality waveform caused by capacitor switching restrike. The waveform is measured at the 28kV bus in a substation. There is only one capacitor on the bus and an oil circuit breaker is used to switch the capacitor.




(a) Voltage waveform




(b) Current waveform

Fig 3-8 Capacitor switching restrike [8]

As Fig 3-8 shows, a significant oscillation appears on the waveform due to the restrike. The resulting distorted waveform might cause further damage. Therefore, the distorted PQ waveform which is related to the capacitor needs to be monitored.

In summary, all the faults mentioned above possess some unique characteristics on their waveforms, which provides opportunities for fault identification. However, as Section 2.4 shows, data-driven methods need to be pre-trained on labelled waveforms before they can be used to identify fault cause. Therefore, the next section will discuss the limitation of labelling and possible solutions.

## 3.4 Labelling of Fault Waveforms

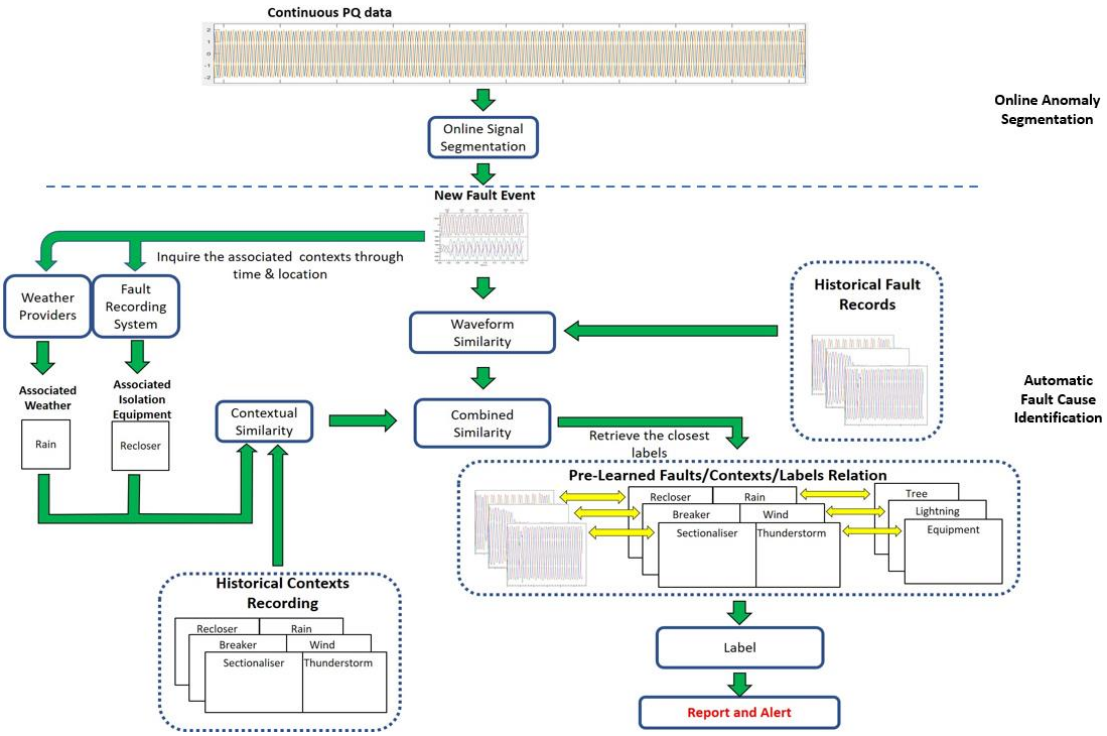
When a newly occurring fault waveform is archived in a repository, some corresponding labels (or contextual data) will also be recorded with it. Common attributes include timestamp, weather, fault phase and isolation equipment. Uncommon but important labels are fault cause and fault location, which usually need a lot of effort to identify. This research is aimed at developing an automated means of fault cause identification. However, as Section 2.4 discussed, limited volumes of labeled waveforms can affect the discriminative power of classifiers. To address this problem, the performance of various classifiers with minimum labeled data can be investigated and the best classifier can be used. However, even collecting the minimal volumes of labeled faults might take excessive time. Therefore, to expedite a fault labelling procedure, additional labeled faults can be automatically generated using other data sources, such as historical maintenance records, to support classifier pre-training. Maintenance reports have been widely used for utilities to schedule and validate remedial works. These are typically free text, which usually contains a large amount of useful fault information, such as fault causes, sequence of protection actions and weather.

The following sections will introduce a novel method which contains these solutions to track the fault causes described in Section 3.3 using the waveform inputs, and the associated labels and maintenance records.

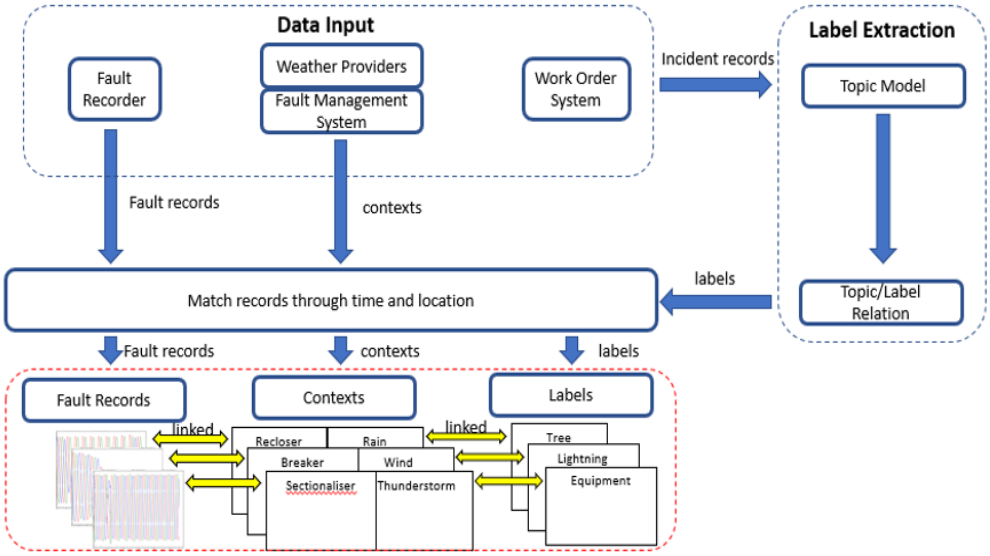
## 3.5 An Architecture for High-Frequency Fault Analysis

As described in Chapter 2, distribution networks are suffering from low observability. To solve this, [1] and [2] proposed a new approach to obtain high-resolution data to find the early signatures of failures and analyse the resulting waveforms to increase situational awareness. This provides more context for remedial and preventive decision support. However, this approach utilised a knowledge-based technique [1][12][13] populated with expert knowledge to analyse faults; developing knowledge-based systems to cover every

eventuality can be time-consuming, and the volume of detected anomalies is excessive [2]. To solve this, a data driven approach is required. However, there are two data-related challenges obstructing the development: the data available to feed-in to the fault detection is excessive in terms of volume, and the historical examples for training fault classifiers are limited. To solve these, some new methods are developed in this thesis for online anomaly segmentation, fault classification, and automatic fault labelling. Fig 3-9 demonstrates an end-to-end framework which uses and integrates these methods.




(a) Online Implementation: Automatic Anomaly Segmentation and Diagnosis




(b) Offline Implementation: Fault Exemplars Generation

Fig 3-9 Proposed automation of fault processing through waveform analysis, (a) Online system: automatic fault segmentation and diagnosis, (b) Offline system: fault exemplars generation

Fundamentally, this system utilises historical anomalous waveforms along with the maintenance reports and the recorded faults context to form an event example repository. The resulting examples can support training of fault classifiers, which is carried out offline as shown in Fig 3-9 (b).

For online implementation in Fig 3-9 (a), when high-frequency continuous waveform signals are streamed in, abnormal events will be automatically segmented out when the transient components are found in waveform which will be specified in Chapter 4, then can be inserted into the fault classifier to interpret the anomalous waveform. While fault data may naturally partition itself according to cause, this is not immediately understandable by a non-expert. To assign a human readable description to faults, the semantics of the closest historical maintenance record are analysed to obtain a label which can then serve as the label for the fault [50]. The specifics of each component will now be elaborated upon.

### *3.5.1 Online System: Automatic Anomaly Segmentation and Diagnosis*

#### *3.5.1.1 Online Anomaly Segmentation*

As Fig 3-9 (a) shows, the approach only uses fault current signals to detect the abnormal events, because current is more sensitive to events. After detection, this event can be segmented then stored in a cache for later processing. Conventional high-resolution anomaly detection schemes, such as using the RMS value of three-phase current to determine abnormal events, can result in high false positive rates [8][6]. This could be resolved through using the shape of the signal as a discriminatory feature to identify the abnormal events [6] rather than the localized thresholding of statistics. Based on the observation in [6], this thesis proposes a new shape-based anomaly detection and segmentation method which can produce the probability of the detection to imply the reliability of the result. The details of this anomaly segmentation model will be described in Chapter 4.

#### *3.5.1.2 Automatic Fault Cause Identification*

Fault recognition amounts to a supervised learning problem, which needs historical examples to train the data-driven model before it can identify faults. To lessen the model's demand for data quantity, a similarity based fault classifier is proposed. If a fault has already occurred, the root cause of the fault can be identified through retrieving the most similar historical event. The similarity measure combines the context of the event with the event waveform. The event context amounts to the associated weather data and isolation equipment record which can be retrieved from weather service providers and online fault recording system indexed on time and location. The details of this fault cause identification model will be discussed in Chapter 5.

### *3.5.2 Offline System: Fault Exemplar Generation*

As Fig 3-9 (b) shows, through aligned time and location records, historical waveforms and contexts can be associated with maintenance reports which correspond to a remedial work order. Maintenance reports have been widely used for utilities to schedule and validate remedial works. Previous research [51] has utilised historical maintenance reports to predict feeder failure and limit the cascading impact of problems, which has validated the use of handwritten records in decision support for the maintenance of the distribution network in New York City.

Fault labelling could be automated by using maintenance reports to create labels for training and validating intelligent classifiers, but these reports are generally not pro-forma based, which results in free-form and sometimes abbreviated text. The labelling model utilises a topic model to learn the relevance between the words from the free text in maintenance reports, then the semantic relevance between documents and topics can be used to categorize the maintenance reports. Topic models are generally based around co-occurring word counts and their resulting statistics, which facilitates generalising their

content into clusters that are considered to be hypothetical topics. The details of fault exemplar generation model will be described in Chapter 6.

## 3.6 Modelling and Simulation to Support Validation of the System

To demonstrate and prove the effectiveness of the proposed fault processing methods and system, the following two approaches are taken: retrospective analysis of archived operational fault data and simulated operational extremes. Archived operational fault data includes DoE/EPRI PQ event repository and two-days continuous data. Among these data, simulated operational extremes and two-days continuous data are used for anomaly segmentation validation, and PQ event repository is used for validating diagnostic function. This section outlines how the simulations were designed, and the nature of the archived data. All of this is used in the future chapters to validate the operation and performance of the methods.

### *3.6.1 Simulated Operational Extremes*

Various faults are simulated to validate the capability of the anomaly segmentation. The severity, duration and location of the faults can be arbitrarily defined in a realistic range. The following two subsections will discuss the modelling detail.

#### *3.6.1.1 Modelling PQ Disturbances*

As noted in [6], power quality measurements under normal conditions on an operational network can be decomposed into noise components and normal load components. Transient events contain additional components which can indicate the operational context of the distribution network [52]. For the waveform, the voltage and current can usually be expressed as:

$$S(t) = \sum_{k=0}^{K} A_k \cos(2\pi k f_r t + \varphi_k) + n(t) + a(t) \tag{3.1}$$

Where $t$ is the time, $S(t)$ is the signal measured at substation, $A_k$ and $\varphi_k$ represent amplitude and phase of $k$th harmonic component, *K* is the highest harmonic order, $f_r$ is the fundamental operation frequency, $n(t)$ is high-frequency random noise and $a(t)$ represents the transient components. In order to rigorously test the capabilities of an anomaly detector, it is necessary to identify the features of PQ disturbances that it may encounter. The following subsections cover the individual aspects of this and discusses how these can be realistically modelled.

#### *3.6.1.1.1 PQ Transients*

Many faults have precursors that are observed long before they evolve into a serious outage. These precursors, such as effects from electrical arcing and mechanical aging, can repeat and generally increase in severity resulting in outright failure of components. The duration of these transients varies, but incipient faults can occur with increasing frequency as they develop further. The duration of the incipient fault development period offers a possibility for utilities to forestall serious faults or outages in advance. However, not all transient components are attributable to faults, some of them are caused by regular operation of the network, such as load switching [53]. These regular events can result in some small changes in signals which might confuse an anomaly detector into raising a false alarm. To address this, some different transient components are simulated to test the robustness of anomaly detection methods. Some transients are associated with faults, such as Kizilcay's arcing model [54] [55], capacitor switching model and three-phase fault model [6]. Others are not, such as load switching model [6]. The underlying methodology for these models is demonstrated in Fig 3-10.

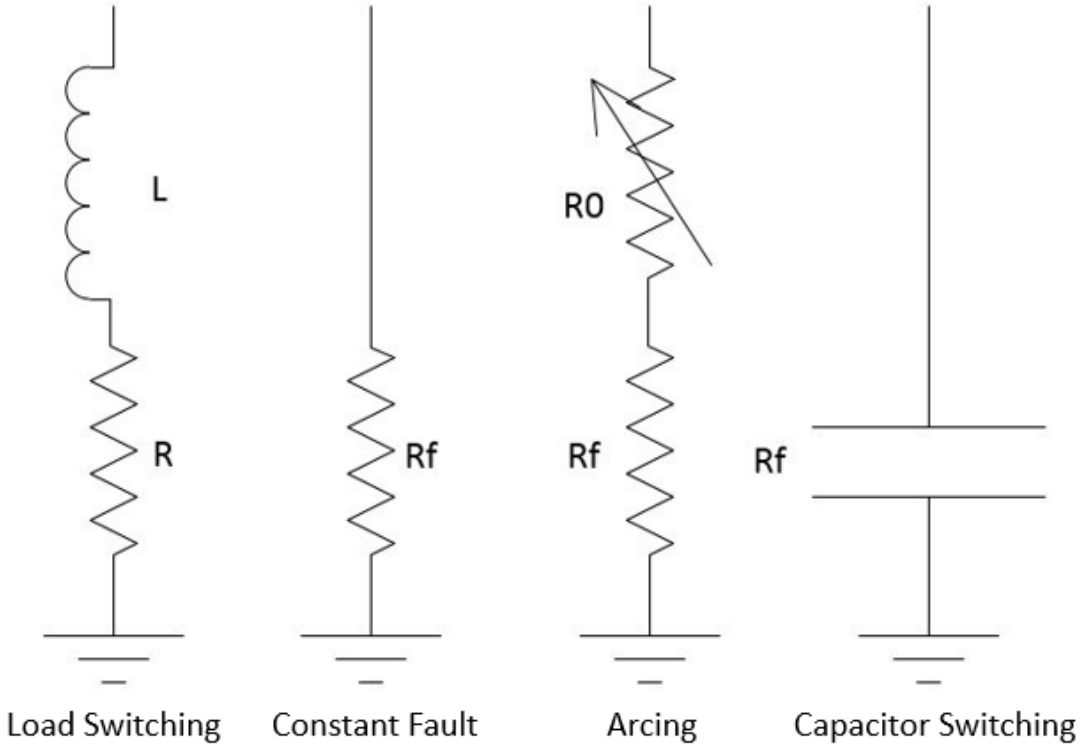



Fig. 3-10. Definition of transient models in terms of RLC

As Fig 3-10 shows, the four models are RLC circuits, which consists of a resistor (R), an inductor (L) or a capacitor (C). These electrical components can be connected in series or parallel. RLC circuit can used to represent different loads or faults. Among these models, constant fault is one of the most common models for anomaly/fault detection. Kizilcay's arcing model has previously been validated [55] and used to test distribution network PQ anomaly detection [6][54]. This model is representative of the arcs in air in a resonant-earthed medium voltage system [55] and also emulates High Impedance Faults (HIF) in overhead lines on distribution networks [55]. These arise in the operational context in two situations: broken overhead lines and faults that earth through arcing suppression coils. In both of these cases the high impedance nature of the fault results in a very small fault current which can be challenging to detect as an anomaly. As Fig 3-10 shows, load switching in the thesis represents the inductive load switching, which is common in daily

load schedules to achieve system balance on a distribution network. Load switching usually results in small transient components in the PQ waveform which could confuse an anomaly detector: capacitor switching is used to correct voltage and power factor, in doing so though, it can bring high magnitude and high frequency transient overvoltages which may damage customer side equipment, resulting from oversensitive tripping in response to these momentary excursions, such as a capacitor switching restrike [56]. It also significantly affects the network health compared with load switching [56]. Although capacitor switching transients can be mitigated by harmonic limiters in the operational context [57], it would be too complex to emulate all of these limiters in the simulation; therefore, it is worthwhile to just model capacitor switching in the simulation network and ignore the unrealistically significant power quality distortion during these events. In summary, for anomaly detection, capacitor switching should be considered as a fault type, but load switching should be treated as a normal event.

#### *3.6.1.1.2 PQ Noise*

PQ noise challenges the correct identification of transient components in anomaly detection [58]. Noise is commonly modelled as high-frequency white noise which is usually caused by Electromagnetic Interference (EMI). EMI can damage the devices and degrade service levels. Conventionally, standards with respect to Electromagnetic Compatibility (EMC) [59] are made to limit the emitted EMI level. Although noise such as EMI can be limited, PQ noise is inevitable, and they can exist on any device, especially for electronic devices [59]. Therefore, it is necessary to verify the robustness of an anomaly detector in the presence of such forms of noise. [6] has verified that PQ noise under normal operation almost obeys a Gaussian distribution.

#### *3.6.1.1.3 Harmonic Content*

PQ harmonics are usually produced by industrial plant (diodes, transistors, rectifier control and frequency convertor, etc.) and power electronic devices in commercial

buildings (computer, monitors and telecom system, etc. ) [59]. This can affect an anomaly detectors performance. Although a lot of work has already proposed some schemes to reduce or eliminate harmonics, high-frequency harmonics can be difficult to completely filter out. Therefore, anomaly detection on a distribution network must be robust to the harmonics from realistic power electronic device penetrations. Furthermore, with a large number of residential customers produce harmonics through domestic PV inverters which can continually affect the network over the long-term, and these can be represented using non-linear load models [60]. To mimic harmonics from a non-linear load, a simple but generic power electronic model [61] [62] is utilised which is demonstrated in Fig 3-11. This model utilises diode switching along with an impedance model to emulate a variety of non-linear loads, such as PCs and Electric Vehicles. Switches, such as diodes, can produce ripples which dominate the harmonics generated by power electronics [60].

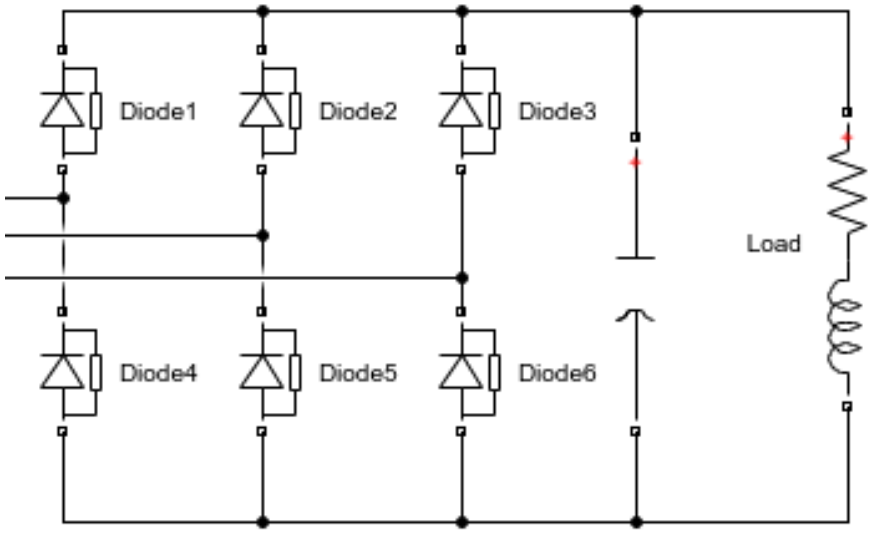



Fig. 3-11. Circuit representation of harmonic model used

#### *3.6.1.2 Modelling the Operational Network*

The IEEE 13-Bus Test Network [63] is an unbalanced radial distribution network at medium voltage level which consists of underground cables, overhead lines, shunt capacitors and a transformer. The injected faults provide the ground truth for abnormal

events which can evaluate whether an anomaly segmentation detector with high-frequency data acquisition is capable of finding anomalies and how early it can detect the event. The network and disturbance configuration are shown in Fig 3-12. According to [6][54], the initial parameter values for the IEEE 13-bus simulation are given in Table 3.1. Using this model, simulation is used to generate events to cover all possible parameter values: 1632 load switching events, 1278 high current arcing faults, 928 low current arcing faults, 480 capacitor switching operations and 560 constant impedance faults. To add additional operational realism, simulation of 10-50A low current arcing is also included which aligns with High Impedance Faults seen in practice [64]. Moreover, the fault duration is randomly assigned within a realistic range, which is generated by a random seed. Except that, the system short-circuit level can affect the disturbance PQ waveform characteristic obtained at the substation. A more robust network will obscure or minimize the changes to PQ waveform. The total load on the network is approximately 3.8MVA. Accordingly, the short-circuit level of the network can be set to 20MVA, 60MVA and 100MVA to test accuracy under different short-circuit levels [65]. Furthermore, some subtle disturbances, such as partial discharge, are difficult to detect because the noise and harmonics conceal the transient components. Through observing the level of white noise and harmonics in realistic data [66], an appropriate level of white noise and harmonics from non-linear loads are added to the simulated signals to validate robustness to noise: 30dB signal-to-noise ratio (SNR) for current and 40dB SNR for voltage are commonly used for power delivery in distribution networks; According to IEEE 519-1992 [67], THD(%) limits of voltage and current are 5% and 15% respectively. This range determines the configuration of the harmonics generated by the impedance model described in Section 3.6.1.3, which is demonstrated in Table 3.2.

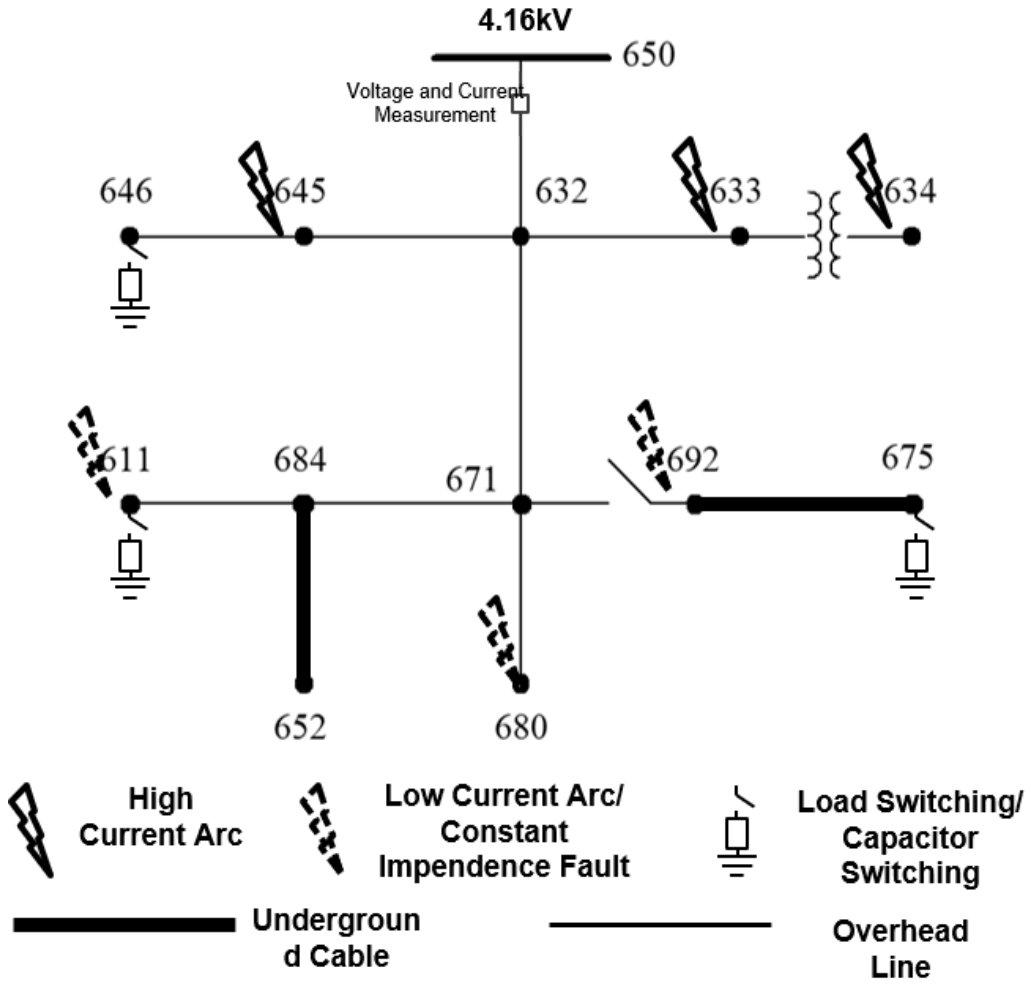



Fig. 3-12 IEEE 13-bus simulation with fault injection

Table 3.1 Fault setup in IEEE 13-bus Simulation [6][54]

| Fault type | Fault Location | Duration (cycle) | Impedance |
|---|---|---|---|
| High Current Arcing Fault | 633,634,645 | 0.25 to 1.25 | u0=300~4000 V<br>r0=0~0.015 ohm<br>tau=5e-5~4e-4 s |
| Low Current Arcing Fault | 680,652,692 | 0.25 to 1.25 | u0=1000~6000 V<br>r0=0.1~100 ohm<br>tau=5e-5~4e-4 s |
| Constant Impedance Fault | 680,652,692 | 2 to 3.25 | Rg=0.1~1500 ohm<br>Rf=0.1~10 ohm |
| Capacitor Switching | 680,652,692 | N/A | C=1e-6~4e-6 F |
| Load Switching | 675,611,646 | N/A | Rg=0.3~10 ohm |

TABLE 3.2 Noise Configuration in IEEE 13-bus Simulation

| Noise type | Fault Location | Duration (cycle) | Configuration |
|---|---|---|---|
| White Noise | N/A | N/A | SNR = 70,80,90 dB |
| Harmonics | 671 | N/A | R=0.1~100 ohm<br>L=0.0005 H<br>C=60e-6 F |

In general, this simulation contains a large amount various faults in a representative radial distribution network. The generated waveform at substation will be used to validate the effectiveness of continuous anomaly segmentation function.

### *3.6.2 Archived Operational PQ Data*

Archived operational faults consist of two databases: a PQ event repository and two-days of continuous PQ data. The PQ Event Repository [46] archives 334 anomalies and faults by US Department of Energy from January 2015 to September 2017 which provides an opportunity to test fault diagnostics and fault labelling on operational faults. Two days of continuous operational 25kV substation data [66] is provided by the IEEE Power Quality Data Analysis Working Group which can be used to test high-frequency anomaly segmentation.

#### *3.6.2.1 PQ Event Repository*

The repository provides the fault waveform and the associated time, weather, fault cause and corresponding field crew fault records. However, it is difficult for a realistic fault repository to contain all of these, especially the fault cause labels - this is because fault cause identification is still a problem for DNOs to identify, none of the standards are made for DNO to record faults information - operational practice tends to be to use free text to record fault case information, there is no standard across DNOs for a pro-forma approach to capture fault details. The sampling frequency of waveforms is 0.96kHz and 3.84kHz –

the PQ events in this repository are collected from multiple locations and multiple devices, which results in the sampling frequency being different. A high-resolution representation is shown in Fig 3-13.

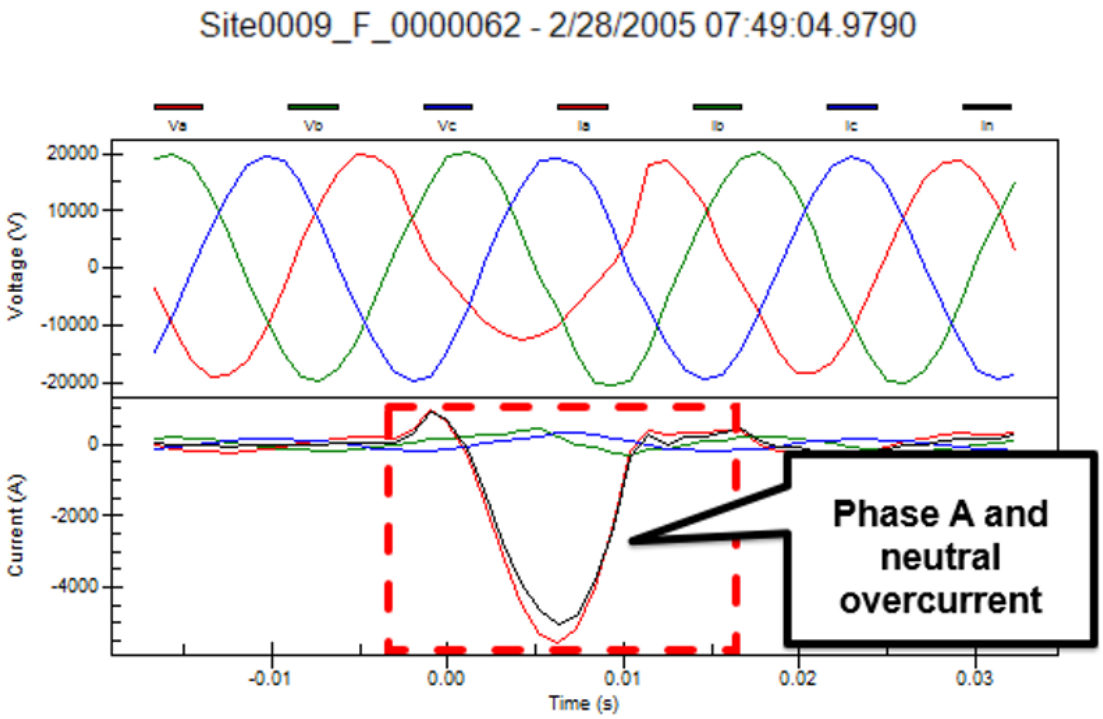



Fig 3-13 PQ waveform representation

As fig 3-13 shows, the library also provides the fault waveform start time and end time down to millisecond level. Table 3.3 shows the additional fault details that accompany each record for the first 14 instances.

Table 3.3 Maintenance records, fault cause labels and associated weather

| EventId | Cause | Weather | Isolated Equipment | Maintenance Report (free text) |
|---|---|---|---|---|
| 0001 | Tree | Clear Weather | Recloser | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 0004 | Tree | Clear Weather | Recloser | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 0005 | Tree | Clear Weather | Recloser | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 0007 | Tree | Clear Weather | Recloser | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 3042 | Equipment | Unknown | Equipment | Equipment, Device UG, Damaged. |
| 0021 | Equipment | Clear Weather | Equipment | Overhead Insulator Failure. BROKEN INSULATOR |
| 0022 | Equipment | Clear Weather | Equipment | Overhead Insulator Failure. BROKEN INSULATOR |
| 0062 | Undetermined | Raining | Undetermined | Storm |
| 0064 | Undetermined | Raining | Undetermined | Storm |
| 0067 | Tree | Thunderstorm | Tree | Tree/Limb Growth |
| 0065 | Tree | Thunderstorm | Tree | Tree/Limb Growth |
| 0068 | Tree | Clear Weather | Tree | VINES ON TRANSFORMER |
| 2760 | Unknown | Unknown | Unknown | Short duration variation. No outage information found. |
| 3048 | Equipment | Unknown | Equipment | Equipment, Capacitor Station, Damaged. |

Table 3.3 shows the free text report associated with the fault indicating its context. Furthermore, Table 3.3 also contains the associated weather data and isolation equipment which can be used as an input for fault diagnosis as Fig 3-9 (a) shows. These data provide ground truths for validating fault recognition and fault labelling. To ensure that the result is not affected by unlabelled data and individual extreme cases, appropriate labelled faults

should be selected. Firstly, the repository has 12 fault causes labels for the fault as Fig 3-14 shows. However, a third of the archive are unknown which are labelled as 'Unknown', 'Undetermined', 'Other' or without any labels which should be removed. This also implies present recording systems will still fail to identify the fault cause for some events. The prevalence of fault cause, as shown in Fig 3-14, is unbalanced which can be challenging for automated classification – it will be difficult to learn general representations of seldom seen events. Five fault categories have sufficient prevalence to be considered: Equipment Caused Faults (ECF), Lightning Strike Faults (LSF), Vehicle Impact Faults (VIF), Animal Contact Faults (ACF) and Tree Contact Faults (TCF). These five categories will be used to test the proposed online fault classification and the offline fault labelling.

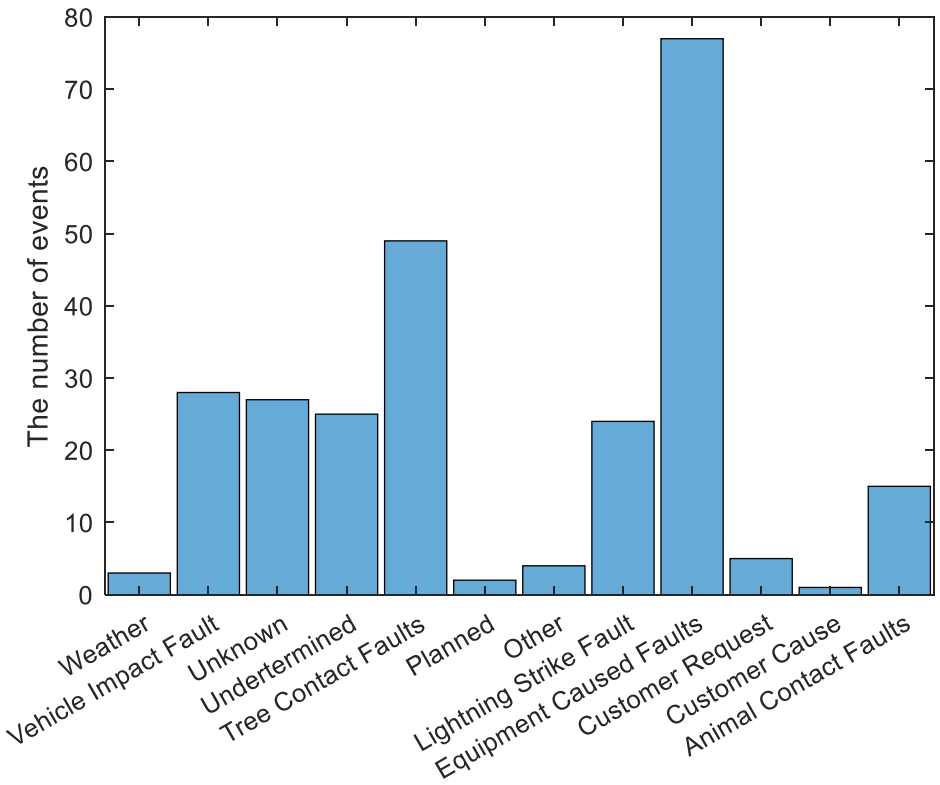



Fig 3-14 DoE repository fault prevalence

### *3.6.2.2 Two-days continuous PQ data*

A realistic sample of two-days continuous PQ data is used to validate the effectiveness of the diagnostic function. As Fig 3-15 shows, the two-days of continuous PQ data is recorded at the substation bus. The topology of this network implies the absence of power electronic devices which may present less of a challenge than the simulation cases presented here; however, more realistic noise make this case still worthwhile to test the detector. Fault occurrences have not been marked up on this data set so prior to usage, anomalous regions are manually identified for evaluating the effectiveness of changepoint detectors, the RMS value for every cycle is extracted and shown in Fig 3-16. According to the RMS value, only a single anomalous region can be found approximately 41 hours into the data, where four abnormal events are picked up. Furthermore, the sampling frequency of the data is 3.84 kHz. This data is used to validate the effectiveness of the online anomaly segmentation function.

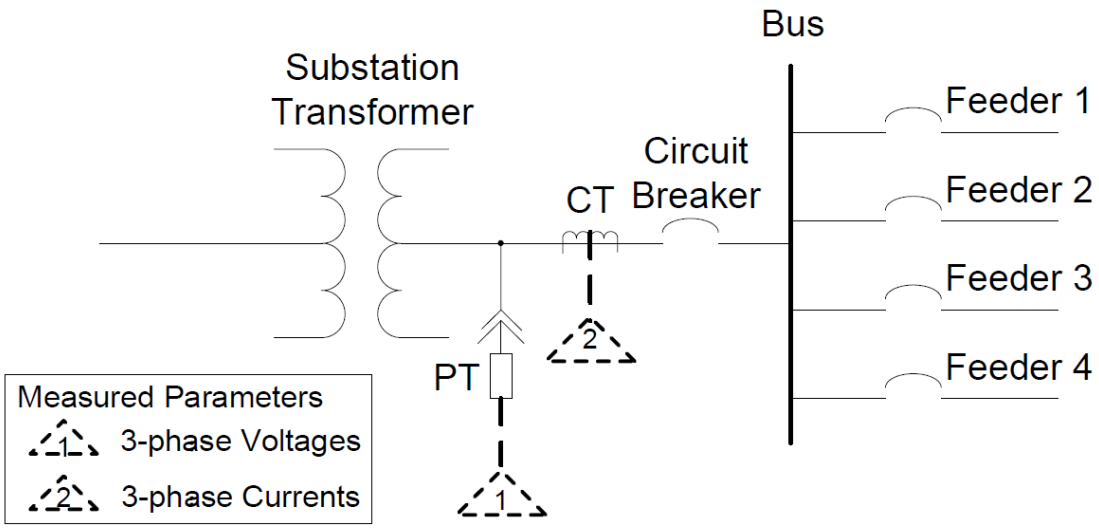



Fig 3-15 The measurement setup for 2-days continuous data [66]

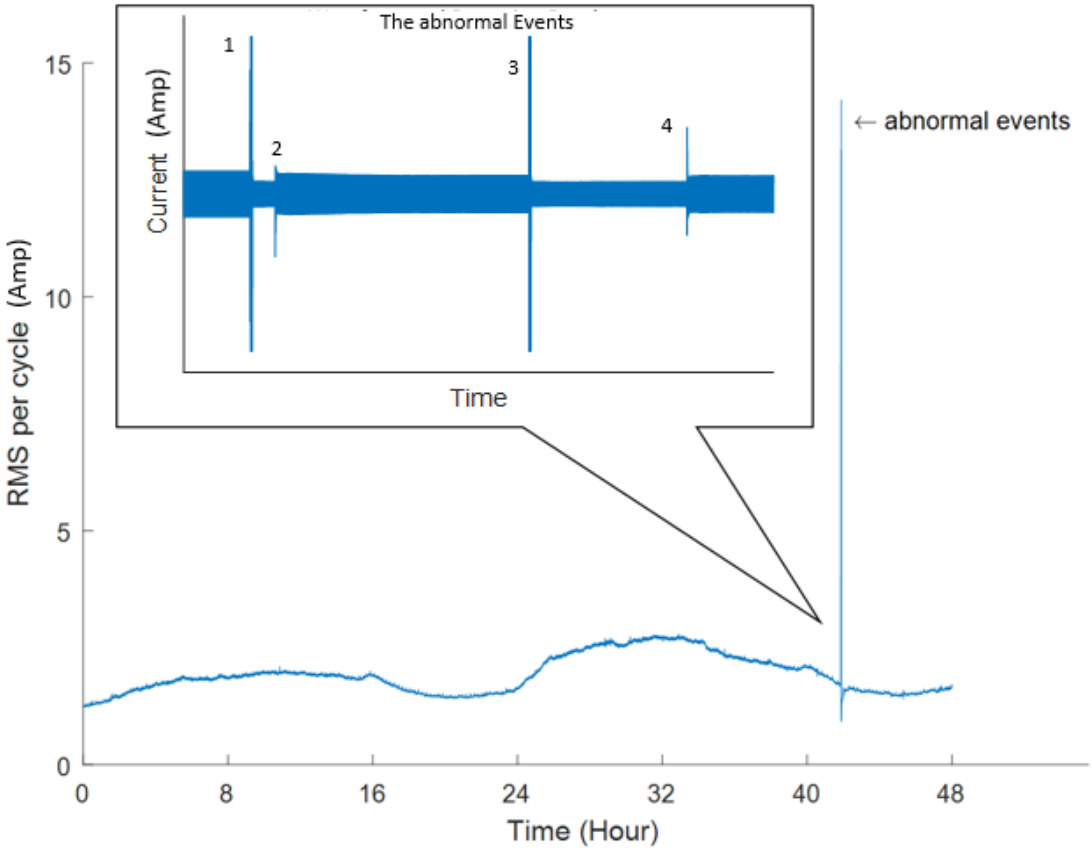



Fig 3-16 anomalous regions for two-day continuous data

## 3.7 Conclusion

Automatic fault cause identification using high-frequency PQ data can help DNOs reduce the occurrence of failures and expedite the corresponding maintenance. This chapter has outlined an end-to-end design for automatic high-resolution fault diagnosis, which includes anomaly segmentation, fault cause diagnosis and automated fault event labelling. The following chapters will expand on these and will use both simulated extremes and archived operational data to test their effectiveness.

# Chapter 4

# Online Power Quality Waveform Anomaly Segmentation for Distribution Networks

## 4.1 Introduction

The previous chapters have discussed that high frequency fault analysis on power distribution networks can help DNOs to achieve higher situational awareness. However, one of the challenges of high frequency fault analysis is to precisely extract anomalous regions from multiple continuous data streams before classifying the underlying fault signature. Failure to do this correctly can result in the needless storage of large volumes of data, which must then be analysed to assess their usefulness. Conventional high-resolution anomaly detection has previously utilised the differential waveform RMS method [48], Mean Absolute Variation in Squared Amplitude (MAVSA) [48] and Kalman Filtering (KF) [68] as examples. However, the former two approaches are prone to false alarms because they are built on local statistical characteristics, such as waveform RMS and mean value, rather than the entire waveform. In other works, [6] proposed a generic anomaly detection method using hypothesis testing to examine the statistic distribution of the PQ noise of a normal waveform, then Kullback-Leibler Distance (KLD) is utilised to determine whether the noise from a new event conforms to the distribution of that for normal operational waveform – a standard Gaussian distribution is used in [6]. None of these works discussed anomaly detection performance in a non-linear noise environment and tested for the online streaming data. With the increasing use of Low Carbon Technologies (LCT) in distribution networks, the harmonics and noise produced by convertors/inverters can degrade the PQ of the grid, and can be challenging to the robustness of anomaly detection [69]. Additionally, previous works have tended to require adjustment of the threshold of anomalous behaviour to accommodate dynamic noise levels, which can also obstruct practical implementation [70]: setting a threshold manually can be time-consuming and prone to error in the presence of previously unseen noise. To solve these, this chapter develops an online and generic PQ anomaly detection method for distribution level faults which can fit in the processing in Fig 3-9 of Chapter 3. The proposed anomaly detection model tackles the complex problem of segmentation of the high-resolution signals and

automatically learns the bounds of anomalous behaviour in an online mode without historical data. This overcomes some of the key barriers and problems outlined in the literature reviewed, and allows a move to effective online anomaly and fault detection.

### 4.2 Automating Disturbance Detection

This section will detail the approach to extracting PQ events from a continuous stream of current sampled at a high resolution. Plant failure on power distribution networks usually starts with PQ distortion, which manifests as either an immediate interruption (e.g. equipment damage, animal contact and lightning strikes) or periodic anomalies (e.g. arcing and aging [71][72]) of changing magnitudes.

The problem is constrained by the following practical considerations:

- **Temporal dependency**: PQ data is non-stationary over time during disturbances and faults which requires the subsequent predictions to depend on the previous observations rather than assuming every observation is independent
- **Weak delineation**: Faults may cause anomalies to be dispersed across time which dispels the assumption that a new underlying regime delineates a fault with a single structural break in stochastic behaviour.
- **Quantification of Uncertainty**: power utilities will need to know how confident they can be in automated decisions made so any approach must offer transparency as to the certainty that a segment of data is anomalous.
- **Online Operation**: The rate at which data is acquired means that online analysis must be conducted as data storage requirements will not be scalable to a network wide solution. Furthermore, online learning of anomalous behaviour is preferable owing to the individual operating characteristics of networks.
- **Adaptive thresholds**: The heterogeneity of anomalies in distribution networks means fixed thresholds will not be sufficient for detection – changes in loading, which occur over longer timescales than faults, render these ineffective.

Based on these requirements, a changepoint analysis approach can be seen as an appropriate solution. Changepoint analysis finds abrupt changes [73][74] in time series data, which can in practice distinguish between different operating regimes, such as identifying industrial plant faults [75]. The conventional changepoint approaches include Hidden Markov Model [76], state space models such as the Kalman Filter [68] and the sequential application of hypothesis testing [77]. The approaches are designed to identify a particular abrupt change; however, individually they do not meet the practical constraints given above. Thus, the approach proposed here is the application of an online Bayesian Changepoint detector [73], which learns online and models the uncertainty of anomaly occurrence across a detection window.

### *4.2.1 Processing Design*

As Fig 4-3 shows, a sampling window roughly segments raw data, which ensures that the abnormal condition is learned only from local transients rather than long-term load variation and reduces the computational complexity. Then an abnormal component extraction algorithm is applied to transform the sinusoid in the window into a transient signal. Following this, the resulting signals are input to a Bayesian Changepoint PQ anomaly Segmentation (BCPQS) model with initial conditions that can be either randomly selected or set using prior knowledge. These stages will now be described in detail.

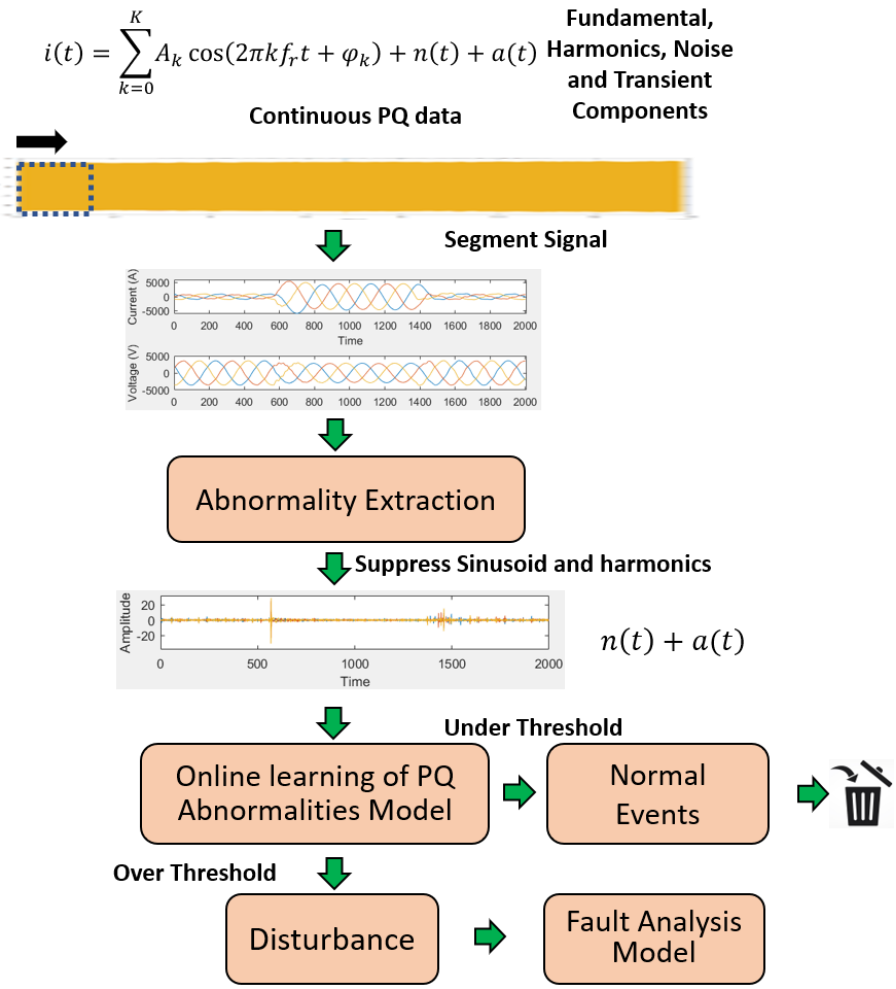



Fig. 4-1. The design of automatic segmenting steaming PQ data

### *4.2.2 Sampling Window Design*

Based on the findings of past work [70], the length of the sampling window is specified as 8 cycles. To ensure the detection result is not affected by the initialization stage at every window, the detection avoids this and double counting anomalies by refreshing the raw data for 7.8 cycles time (i.e. less than window length) to capture the continuous input signal.

### *4.2.3 Abnormality Extraction*

As described in [6], the current signal can be thought of as a sinusoid (the AC fundamental as well as harmonics) with abnormalities (noise and transient components) superimposed. The conventional approach to decoupling the sinusoidal components from

the abnormal components of the signal are to superimpose the faults over the previous cycle [48]; this approach is adopted here to reduce the computational complexity and accommodate the supposed absence of knowledge of the form of faults. This method utilises the present measure superimpose over the last health cycle which can be simply expresses as:

$$n(t) + a(t) = i(t) - i(t - qN_0\Delta t) \tag{4.1}$$

Where $n(t)$ and $a(t)$ refer to noise component and transient component to maintain consistency with the Equation (3.1); $q$ denotes the number of the gap cycles between last healthy cycle and the currently measured cycle. $N_0$ denotes the number of samples in one cycle current; $\Delta t$ is the sampling interval which represents the time gap between two consecutive samples. In short term, the fundamental component and harmonics would not change; therefore, the superimpose would only reveal the noise and transient components.

The extracted component residual has been validated as Gaussian distributed for normal operation using a normality test [6]. To investigate the characteristics of the noise of the 2-days of continuous PQ data, 2400 normal operation signal pieces (50 events per hour) were randomly selected from the archived data to test their Gaussianity using the Anderson-Darling test (AD), Lilliefors test (L), Jarque-Bera test (JB), Shapiro-Wilk test (SW) and Kolmogorov-Smirnov test (KS) [78]. 8 cycles, which amounts to the window size, are used for this Normality test. Comparison of the percentage of the events obeying Gaussian distribution with significance level is shown in Fig 4-2; significance level is the probability of rejecting the null hypothesis when it is true – lower significance level means it is more likely the hypothesis is true (not due to chance) [78].

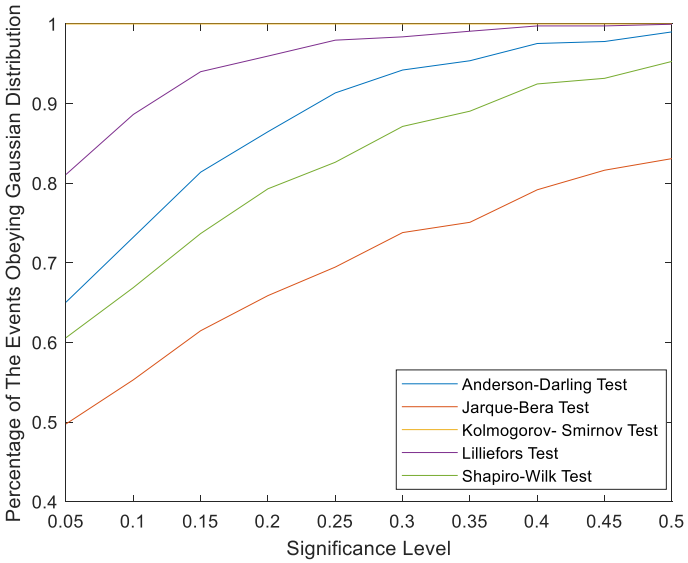



Fig. 4-2. Normality Test for the extracted noise of the 2-days continuous data

As Fig 4-2 shows, KS test and Lilliefors test shows most of the noise obeys a Gaussian distribution, even at low significance levels. AD test, SW test and JB test performs relatively worse, especially for JB. This can be because the latter three tests have a heavier weight for testing tails [78], and the industrial noise has heavier tails than normal Gaussian Distribution; one example of industrial noise is visualized using a Quantile-Quantile (Q-Q) plot [79] in Fig. 4-3.

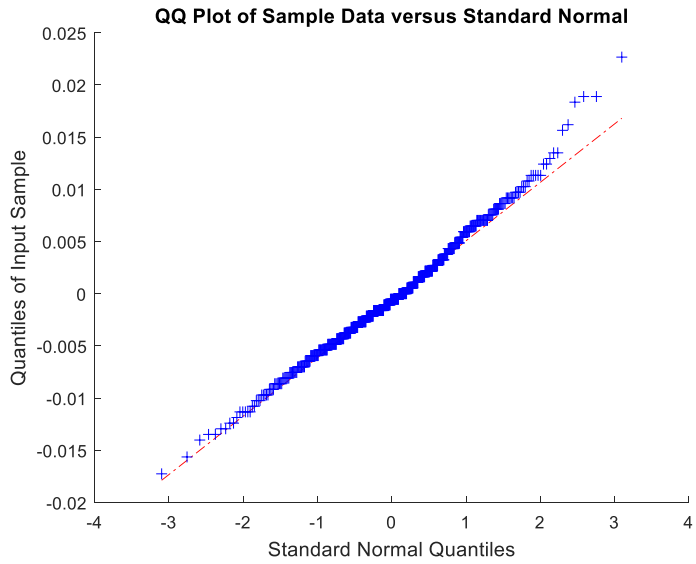



Fig. 4-3. QQplot for one of the extracted noise from the 2-days of continuous data

As Fig 4-3 shows, the dash line represents the standard Gaussian distribution, and the blue data represents the selected signal. The distribution of the noise is close to the shape of a Gaussian distribution, but it has heavier tails than a Gaussian distribution. Therefore, the Gaussian distribution is not a best distribution to track the noise. In following sections, a new distribution to track the noise will be introduced.

### *4.2.4 Online Learning of PQ Abnormalities Model*

After abnormality extraction, the processed signal of the single phase current is then run through a changepoint detection model [73] to identify abnormal events. To identify changepoints within the sampling window (given in Section 4.2.2), the problem can be formulated as $P(S_i|x_{1:i})$, where $i$ is the time step, $x_1 \cdots x_i \cdots x_k$ are a sequence of PQ current observations, $x_k$ is the last observation in the sampling window and $S_i$ is the binary state (healthy or abnormal) at time $i$. Thus, the formula represents a probability of the normal or abnormal state at time $i$. If the probability strongly indicates an anomaly existed at time $i$, then the waveform at that time could be saved and put in a fault analysis model for further processing. The process of online segmentation of a PQ abnormality for a single detection window is shown in Fig 4-4. Online learning of the bounds of an anomalous region works on the principle of run length estimation [73]. that is, how long a model can be sequentially updated with new PQ observations before a random model of behavior is considered to be more likely than the current one. To achieve this in this application context, BCPQS assumes the parameters of the start point in the sampling window have been reset to represent a normal event – this is undertaken offline using historical normal events. Each new observation updates model hyper-parameters, which are then utilised to evaluate the changepoint probability and run length distribution. Finally, the run length probability is normalised and used to detect anomalies by comparison with probability thresholds learned offline. When a changepoint has been encountered, the prior distribution is reset, and parameters are sequentially calculated as before until another

changepoint is found. After a start point is found, the fault segmentation function is triggered to look for the corresponding end point of the abnormal event and segment it out. The end point is determined when three consecutive cycles are free of changepoints [70].

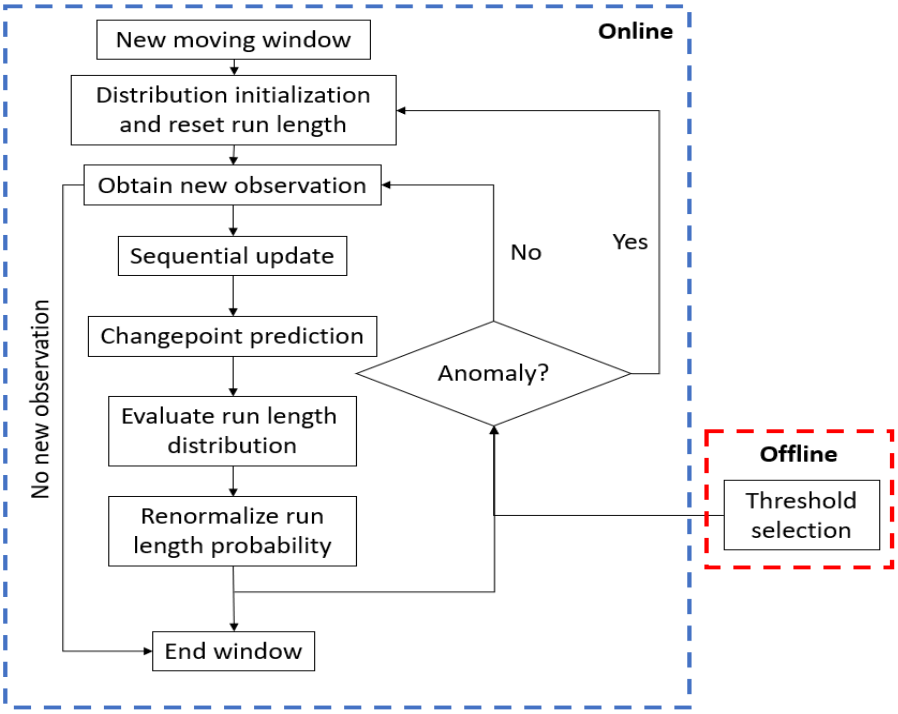



Fig. 4-4 Bayesian Changepoint anomaly detection workflow process

This section will now detail: the initialization of the BCPQS model parameters, how these are sequentially updated as more data is observed and how the detector model works in the presence of PQ noise.

#### *4.2.4.1 Noise Distribution Initialization*

As discussed in [6], the non-periodic component of a normal signal's behaviour can be modelled by a Gaussian distribution with the presence of any abnormalities resulting in the violation of this assumption. A key part of BCPQS arises from the adoption of the exponential family form [80] of a Gaussian distribution with conjugate priors over its parameters. Conjugate prior is defined as a prior distribution, which is from the same

distribution family as the likelihood and posterior [80]. The Bayesian formulation of this results in a Student-t distribution being used to track abnormal performance as a probability rather than a Gaussian distribution which has the additional benefit of being more robust to outliers (through its heavy tails) [81]; this can perfectly solve the heavy tails problems described in Section 4.2.3. The distribution is parameterized by prior hyperparameters which in turn can be used to formulate expressions for the mean, variance and degree of freedom for the distribution by integrating over all possible values of the prior distribution[73] .

The conjugate prior of the Student-t distribution is the Normal-Gamma distribution [81], $\mu, \sum \in NG(u, k, \alpha, \beta)$ which is expressed as:

$$NG(\mu, \sum | u, k, \alpha, \beta) = Z \cdot \sum^{\alpha - \frac{1}{2}} \cdot \mathrm{e}^{(-\frac{\sum}{2}(k(\mu - u)^2 + 2\beta))} \tag{4.2}$$

$$Z(u, k, \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma\,(\alpha)} (\frac{2\pi}{k})^{-\frac{1}{2}} \tag{4.3}$$

Where $u, k, \alpha, \beta$ are hyper-parameters and $\Gamma$ denotes the Gamma function. Gamma function is a well-known factorial generalization function amounting to an approximation of a continuous factorial of a real number. The initial hyper-parameters of the observation distribution can change the shape of the distribution which affects the sensitivity of the changepoint detection [81]. The initial values of $\mu_0$ and $k_0$ can be determined by the mean and variance of the mean value of historical normal behaviour data. According to [81], the variance of sample is determined by the initial parameters $\alpha_0$ and $\beta_0$ for the Gamma distribution, All the initial parameters can be estimated from the historical healthy condition data.

#### *4.2.4.2 Sequential Update of Noise Distribution*

As Section 4.2.4 described, the parameters of the predictive distribution need to be sequentially updated based on previous observations. To achieve this, a sequential Bayesian

estimation is utilised to derive parameter updates [81]. Bayesian statistic is a theory to calculate or update probability, which is based on an assumption that the probability can be dependent to the prior knowledge rather than only to the frequency of the occurrence of an event [82]. Based on some pertinent inferences, the corresponding Bayesian posterior [81] is:

$$u_{i+1} = \frac{k_i u_i + \bar{x}}{k_i + 1} \tag{4.4}$$

$$k_{i+1} = k_i + 1 \tag{4.5}$$

$$\alpha_{i+1} = \alpha_i + \frac{1}{2} \tag{4.6}$$

$$\beta_{i+1} = \beta_i + \frac{1}{2}(x_j - \bar{x})^2 + \frac{k_i(\bar{x} - u_i)^2}{2(k_i + 1)} \tag{4.7}$$

This permits a sequential update to be carried out using just the sufficient statistics of the distribution. Sufficient statistics means that no other statistic needs to be calculated from the same sample to provide additional information as to the value of the parameter. In the case of the Gaussian distribution, this amounts to a running sum for the mean. The new parameters are then estimated based on the previous distribution and the current observation. Then the updated new parameters can produce a new Student-t distribution to evaluate the next changepoint probability.

#### *4.2.4.3 Changepoint Detection*

The BCPQS model can be used to evaluate the run length probability of a PQ noise signal over a window. Fig 4-5 shows an example of how the exponential family distribution works on PQ noise and how the model update for new observations. For each run length probability calculation, the hyper-parameters can be updated with each new observation injected, then put in Normal-Gamma distribution to calculate the distribution prediction and run length probability. The run length probability is expressed as:

$$P(r_i, x_{1:i}) = \sum_{r_{i-1}} P(r_{i-1}, x_{1:i-1}) P(r_i | r_{i-1})\, P(x_i | u_i, k_i, \alpha_i, \beta_i) \tag{4.8}$$

Where $r_i$ is run length at time $i$, $r_i = r_{i-1} + 1$ for normal event and $r_i = 0$ for anomaly.

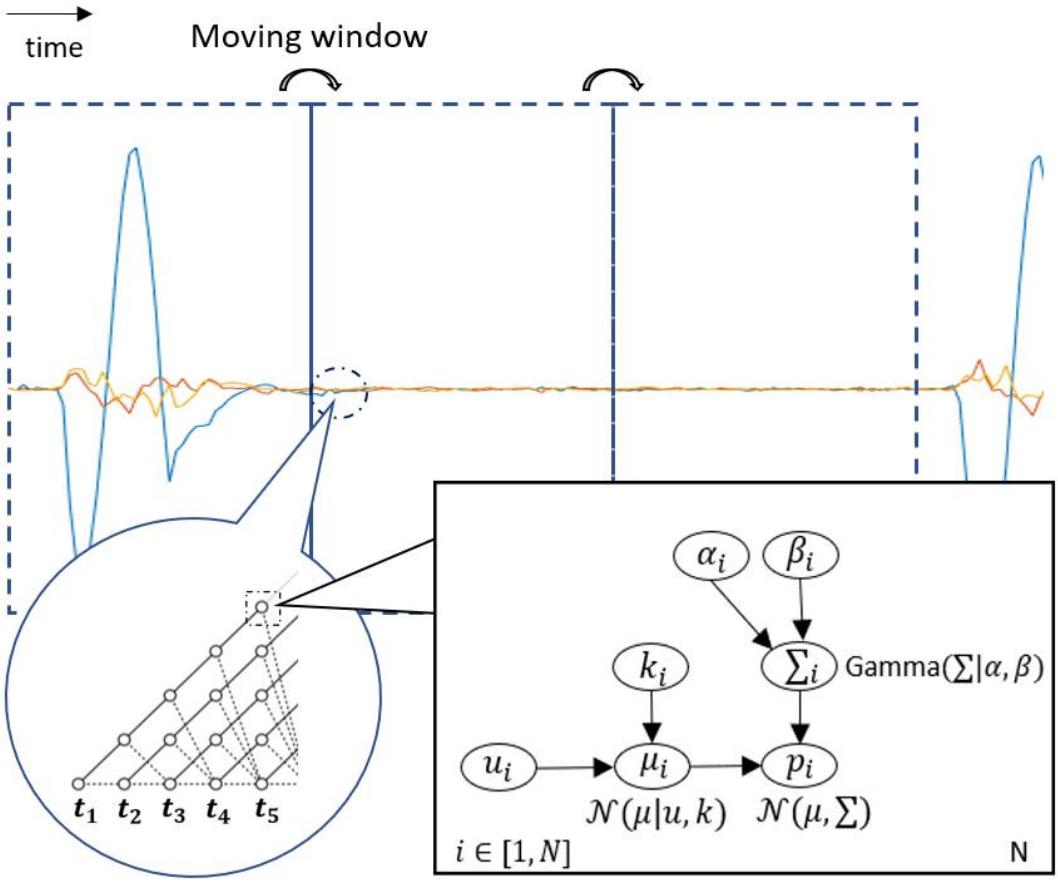



Fig. 4-5 Window based tracking of changes in noise distribution

The run length probability can be calculated for every possible scenario and can form a trellis of probable run lengths over the sampling window.

### *4.2.5 Offline Threshold Selection*

The proposed detector adapts a probability threshold to track anomalies which can achieve adaptive threshold as a certain value can be set for all the cases; however, an appropriate probability threshold and associated approach are required to pick up the abnormal segmentation from continuous current signals. The run length specifies the

number of time points for which the observation distribution has been valid – if this is lower than that for a shorter run length, it is indicative that it should be reset. Owing to the time period required for a fault to develop, $r_i = 5$ has been observed experimentally to identify the minimum plausible run length at a given time. This is because the purpose of the detector is to detect significant disturbances, such as incipient faults, rather than subtle inferences from normal operation, such as demand fluctuation; if $r_i$ is too small, the detector can be too sensitive to the outliers; large $r_i$ makes it possible to pick up changepoints with small variance. An example is demonstrated in Fig 4-6. The run length is sensitive to waveform distortion. As demonstrated in Fig 4-6 (1) and 4-6 (2), the small signal inference around the time 900, 1000, 1140 and 1280 can results in a run length reset but these are not expected to detect. As demonstrated in Fig 4-6 (3), $r_i = 5$ make it more intuitive to detect the start point and end point through setting an appropriate probability threshold.

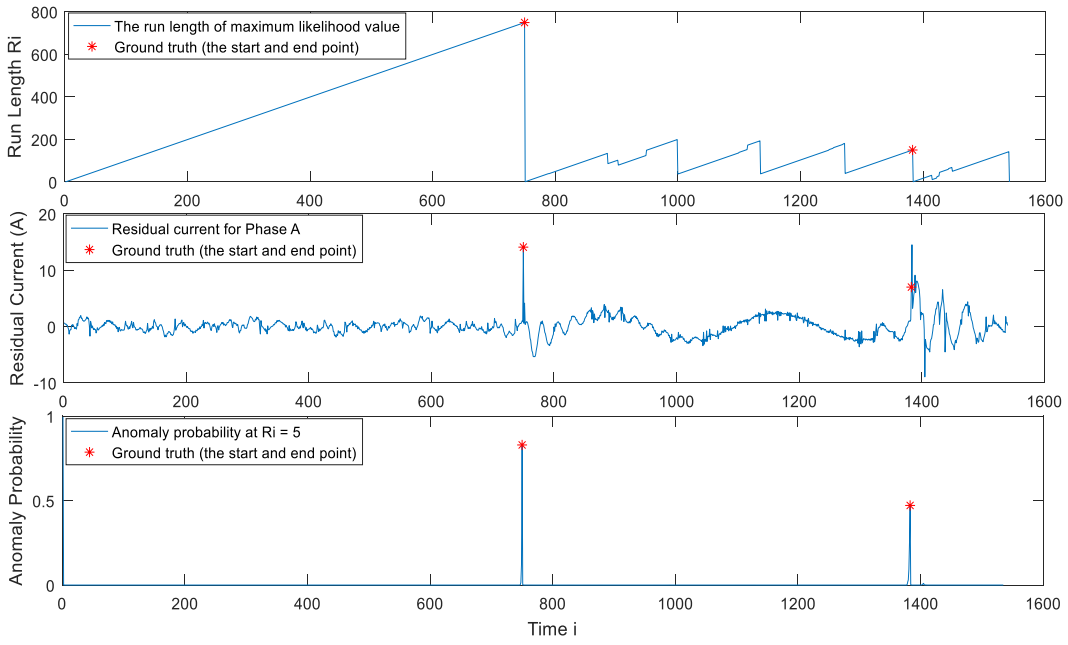



Fig 4-6 Comparison of indicators for changepoint result, (top) the run length of the maximum likelihood varies with time, (centre) residual current varies with time, (bottom) anomaly probability at $r_i = 5$ varies with time

Threshold of anomaly probability is chosen through analysis of historical events and how they affect the false alarm rate [6], and a buffer of three extra cycles on either side of the anomalous region is retained to capture subtle transients as noted in [70].

## 4.3 Operational Case Study

To test the model capability, simulated and operational distribution network monitoring data is used. The purpose of using both is to ensure sufficient operational realism, in the case of the operational data, and the full spectrum of fault possibilities, in the case of the simulation. The BCPQS is now compared against several existing works to demonstrate its effectiveness and practical advantages.

### *4.3.1 Model Initialization*

Section 4.2.4.1 noted that normal operational data is required for the initial values of the prior distributions. In order to achieve this online, the first 200 moving windows, each comprising 8 cycles of normal operation [2], are used to initialize the prior distributions by fitting a Student-t distribution by Maximum Likelihood Estimation. Using the resulting location, shape and degree of freedom estimates for the sample Student-t, the initial values of $u$, $k$, $\alpha$ and $\beta$ can be recovered [30] from:

$$\nu = 2\alpha \tag{4.9}$$

$$\mu = u \tag{4.10}$$

$$\lambda = \sqrt{\frac{\beta}{\alpha} \times \frac{1+k}{k}} \tag{4.11}$$

Where $\nu$ is degree of freedom, $\lambda$ is the scale factor and $\mu$ is the location factor. Sensitivity analysis reveals that the mean of the current almost remains stable at 0 which indicates the initial Gaussian mean can (intuitively for normal operation) be set to 0.

### *4.3.2 Benchmark Changepoint Detection Models*

The Kalman Filter [68], Differential RMS [6][70] and MAVSA [6] detectors are used as benchmarks. Differential RMS detects anomalies through observing whether the differential waveform between consecutive cycles is out with a predefined threshold [6]. Firstly, the differential current waveform $\Delta I$ extract abnormality by superimposing the fault over previous cycle. Then the differential waveform will be divided into $M$ pieces and the RMS value will be calculated for every segment. The normalised segments can be expressed as:

$$\alpha_{l,r} = \frac{\Delta I_{l,r}}{I_r} \tag{4.12}$$

Where $l$ and $r$ denote the $l$-th segment and the $r$-th cycle. The maximum ratio in every cycle is utilised to determine abnormality and the detection rule is:

$$\text{Normal, } \max(\alpha_{l,r}) \leq \alpha_{TH} \tag{4.13}$$

$$\text{Abnormal, } \max(\alpha_{l,r}) > \alpha_{TH} \tag{4.14}$$

Where the $\alpha_{TH}$ is the threshold whose value can be evaluated by using historical data.

MAVSA compares the difference of the squared value between two consecutive cycles with predefined thresholds to detect anomalies [6]. This is similar to the differential RMS, but variance rather than RMS to determine the abnormality.

Anomaly detection with a Kalman Filter utilises the sequential estimation of a non-stationary Gaussian distribution from a dynamic signal; the signal is segmented through comparing the log likelihood of the observation with a predefined threshold – the threshold is calculated at 3.5% false alarm rate using ROC curve [68]. The Kalman Filter can be used to model the evolution of time-varying variables. The state space equations for a Kalman Filter are:

$$q_t = q_{t-1} + w_t \tag{4.15}$$

$$y_t = F_t q_t + v_t \tag{4.16}$$

where $F_t$ are the input time-varying observations, $q_t$ are state variables, $w_t$ is the state noise and $v_t$ is the observation noise; both obey a Gaussian distribution. The parameters are updated using a Kalman filter which has two steps: prediction and correction.

Prediction:

The predicted state variable, $\hat{q}_t$, and predicted covariance, $\widehat{\sum_t}$.

$$\hat{q}_t = q_{t-1} \tag{4.17}$$

$$\widehat{\sum_t} = \sum_{t-1} + \mathrm{W}_t \tag{4.18}$$

$\mathrm{W}_t$ is the state noise covariance at time t.

Correction:

The observation mean and covariance updates are:

$$q_t = \hat{q}_t + \mathrm{K}_t(y_t - F_t\hat{q}_t) \tag{4.19}$$

$$\sum_t = \widehat{\sum_t} - K_t F_t \widehat{\sum_t} \tag{4.20}$$

Where $K_t$ is Kalman Gain which can be expressed as:

$$K_t = \frac{\widehat{\sum_t} F_t^T}{{v_t}^2 + F_t \widehat{\sum_t} F_t^T} \tag{4.21}$$

Then the observation function is expressed as:

$$p(y_t) = \mathcal{N}(\widehat{y_t}, {v_t}^2 + F_t \widehat{\sum_t} F_t^T) \tag{4.22}$$

The abnormal event will be detected only if the evidence (observation probability) of the data point is low, which will be the case if the data at this point contains more noise.

### *4.3.3 Performance Evaluation Criteria*

Continuous streams of monitoring data will mostly represent normal network operation. This resulting high prevalence can result in misleading evaluation metrics which can indicate a very high accuracy even when no detection takes place. However, abnormal events usually should be more important than normal events. This necessitates metrics to reward different scores for True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates and use these to reflect detector performance. Furthermore, a well performing anomaly detector requires correct detection of the abnormal time location: if the error between the detected time and the actual abnormal time is too large, this can result in the method segmenting an incomplete signal out. This can be detrimental to the subsequent fault analysis. Therefore, an evaluation window is required to score every instance. Receiver Operating Characteristic (ROC) is a conventional metric used to evaluate the performance of anomaly detection [6]. The ROC is used to observe the true positive rate (detection probability) versus the false positive rate (false alarm probability) at various threshold setting of anomaly detection. The area under the ROC curve reflects the effectiveness of the detector. However, ROC cannot reward early detection.

The Numenta Anomaly Benchmark (NAB) score [83] has been used to evaluate online anomaly detection against its ground truth. NAB labels ranges of data as anomalous, therefore, it can reward short period error detection and penalize detection outside the scoring window as Fig 10 shows and is defined as:

$$S = \frac{2}{1 + e^{c(y-n)}} - 1 \tag{4.23}$$

Where $S$ represents the performance score, parameter $c$ can control the rate of the drop, parameter $n$ can be used to adjust the zero-cross point and $y$ is the output of

anomaly detection. Accordingly, if a disturbance lasts for three consecutive cycles, the segmentation can be identified as an abnormal event [70]. Therefore, missing detection by more than three cycles should be heavily punished by evaluation metrics as Fig 4-7 shows. To reward a small detection error and heavily penalize missed detection, a narrow reward range is required – outside of the range will be gradually penalized – the accurate detector can be selected, and the curve is expected to decrease more gently at the beginning. Therefore, c=0.02, n=N*1.5 where *N* is the number of samples in one cycle. For multiple detection within a single detection window, only the earliest result is used to contribute to the score. The overall score is:

$$S_t = A_{TP}S_{TP} - A_{FP}S_{FP} - A_{FN}S_{FN} \tag{4.24}$$

$$NAB = \frac{S_t - S_t^{null}}{S_t^{perfect} - S_t^{null}} \tag{4.25}$$

Where $A$ are weights for the confusion matrix elements to raise or lower the importance of true or false positives, and $NAB$ is the ultimate score considered with all instances. $S_t^{perfect}$ represents all the instances are detected with best performance and $S_t^{null}$ means the worst performance for all the instances. As the start of Section 4.3.3 mentioned, True Positives (TP) should be much more important than True Negatives (TN). Therefore, the award of a TP should be higher than that of TN meaning $A_{TP} \gg A_{TN}$. Therefore, 0 is selected for $A_{TN}$ with all others set to 1. The NAB score can obtain 0 for the worst case and 100 for perfect detection. Consequently, a NAB score of over 65 can be considered acceptable [83].

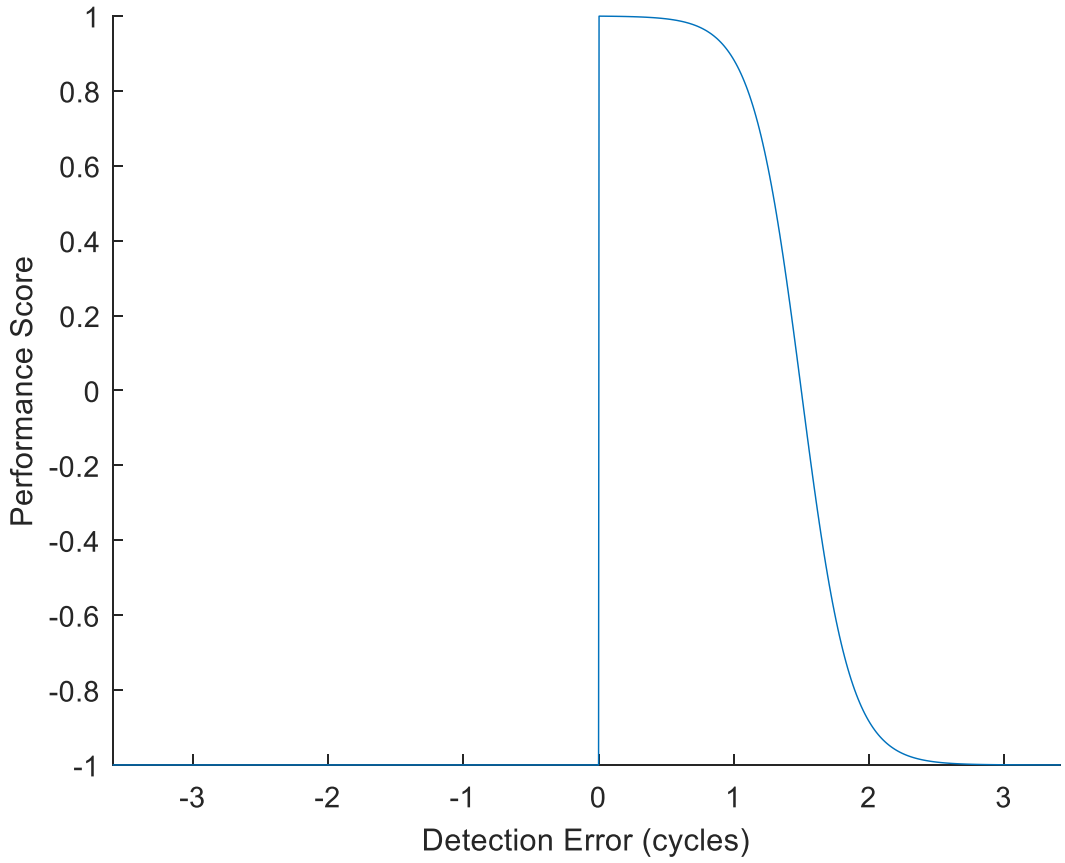



Fig. 4-7 The NAB scoring window shape to visualize the reward zone and penalization zone

## 4.4 Anomaly Segmentation Result

This section will discuss the performance of different detectors using both the simulated and operational data described in Chapter 3.6. Since there are tradeoffs to consider with detection accuracy, a number of metrics are used for evaluation.

### *4.4.1 Detection Performance Using Simulated Network Data*

To test the capability of segmentation, the simulated fault data (discussed in 3.5.1) is used; the data contains waveform data, and the ground truth of start point and end point of faults. In this experiment, the segmented time point will be used to compare with the ground truth, then the detection accuracy taking account of false alert rate is examined by ROC curve (discussed in Section 4.3.3), the detection error for different fault identification,

and the detection error of the start point and the end point are calculated using NAB (discussed in Section 4.3.3). The fault data from Section 3.6.1 is utilised to estimate the ROC curve shown in Fig 4-8.

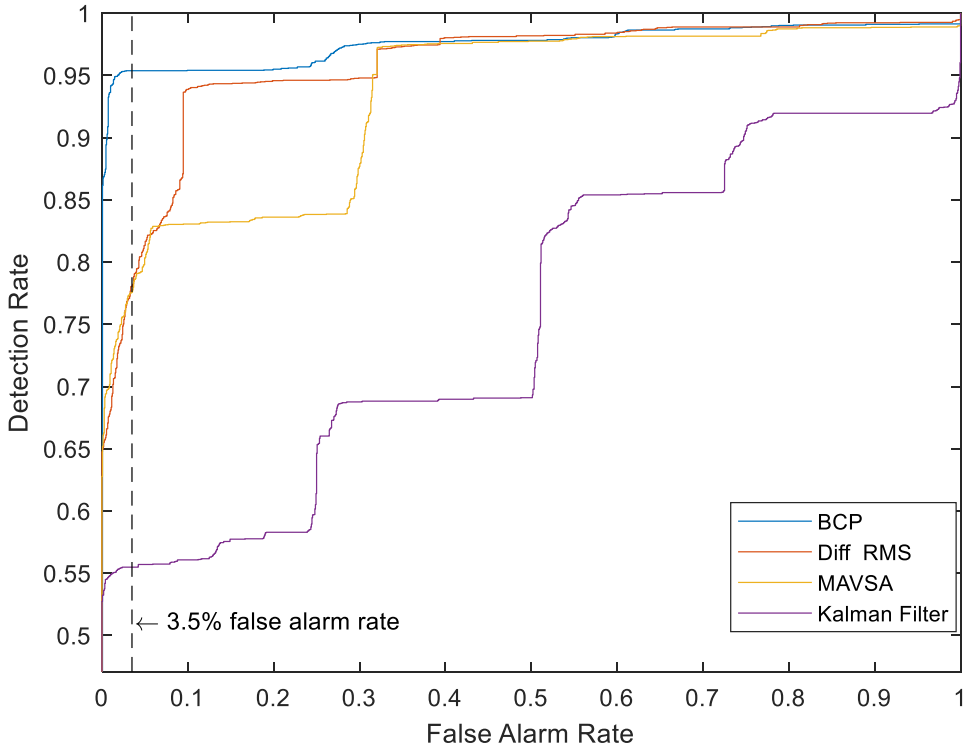



Fig. 4-8 ROC curve of comparing the proposed method against benchmarks

From Fig 4-8, the proposed BCPQS takes larger area of the ROC curve which means it outperforms other detectors for the data with non-linear load noise. Furthermore, a threshold of 3.5% specified by [6] at the same false alarm rate can be calculated for the following performance comparison, and the thresholds are demonstrated in Table 4.1.

BCPQS outperforms other detectors for all the fault recognition as Table 4.2 demonstrates. KF is good at identifying constant impedance, but poor for arcing fault detection. Differential RMS and MAVSA get similar performance, as Table 4.2 shows, and are good at identifying significant over-current change. However, both Differential RMS and MAVSA are built on local statistical characteristics, such as RMS and Variance, which

can result in the false positive rate is too high for practical implementation [70]. Finally, the start point and the end point of the event are used to validate the capability of segmentation, as Table 4.3 shows. Generally, the start point of the abnormal events are easier for detectors to capture which may be caused by the way faults gradually end rather than abruptly change back to normal operation.

Table 4.1 Thresholds of Detectors at 3.5% False Alarm Rate

| Detector | BCPQS | KF | Differential RMS | MAVSA |
|---|---|---|---|---|
| Thresholds | 0.0747 | 91.4024 | 0.0426 | 20963.88 |

Table 4.2 NAB Score of Fault Detector Performance, at False Alarm Rate 3.5%, Starting Time Detection

| Fault type | BCPQS | KF | Differential RMS | MAVSA |
|---|---|---|---|---|
| High Current Arcing Fault | **98.1** | 61.66 | 87.38 | 85.71 |
| Low Current Arcing Fault | **92.15** | 13.50 | 64.38 | 65.63 |
| Constant Impedance Fault | **92.66** | 85.86 | 83.84 | 85.43 |
| Capacitor Switching | **86.68** | 74.22 | 68.99 | 66.60 |
| Load Switching | 96.65 | **97.74** | 96.52 | 96.51 |

Table 4.3 Anomaly Start and End Point Test, NAB Score at False Alarm Rate 3.5%

| Reference Point | BCPQS | KF | Differential RMS | MAVSA |
|---|---|---|---|---|
| Start Point | **94.46** | 53.86 | 74 | 73.79 |
| End Point | **86.65** | 34.45 | 66.24 | 64.23 |

### *4.4.2 Detection Performance Using Operational Data*

Using the set of continuous operational data described in Section 3.6.1.1.1, the proposed BCPQS method is compared against the Differential RMS method with results shown in Table 4.4. Because there are only 4 anomalous episodes in this data, there are too few cases to produce a ROC Curve. This also inhibits the selection of a threshold for the benchmark Differential RMS model: if the same thresholds as Section 4.41 are used to pick up the abnormal events, the proposed method will continue to work well but the benchmark method produces thousands of false alarms – a NAB score of 0.00525. This is because the benchmark methods work on the absolute magnitude of signal which can be sensitive to different voltage and current levels. Therefore, in order to continue to make a comparison the threshold used for the simulation data in Section 4.41 is supplanted with the fixed threshold proposed in [8] for use with Differential RMS – which raises its NAB score to 72.84. Under these conditions, both methods are able detect the four abnormal events with no false alarms. However, the BCPQS model detects the abnormal event earlier than the Differential RMS method as Fig. 4-9 shows. The resulting NAB scores are shown in Table 4.4 where BCPQS is demonstrated as the significantly better performer. This can be attributed to the BCPQS model building a distribution of plausible values of the next value [81] in the PQ current using the values observed since the last changepoint rather than a single prediction being compared to a single threshold.

Table 4.4 NAB Score for Operational Data Anomaly Detection

| NAB | BCPQS | Differential RMS |
|---|---|---|
| Using threshold from simulated data at 3.5% false alarm rate | 95.07 | 0.00525 |
| Using the threshold from c | 95.07 | 72.84 |

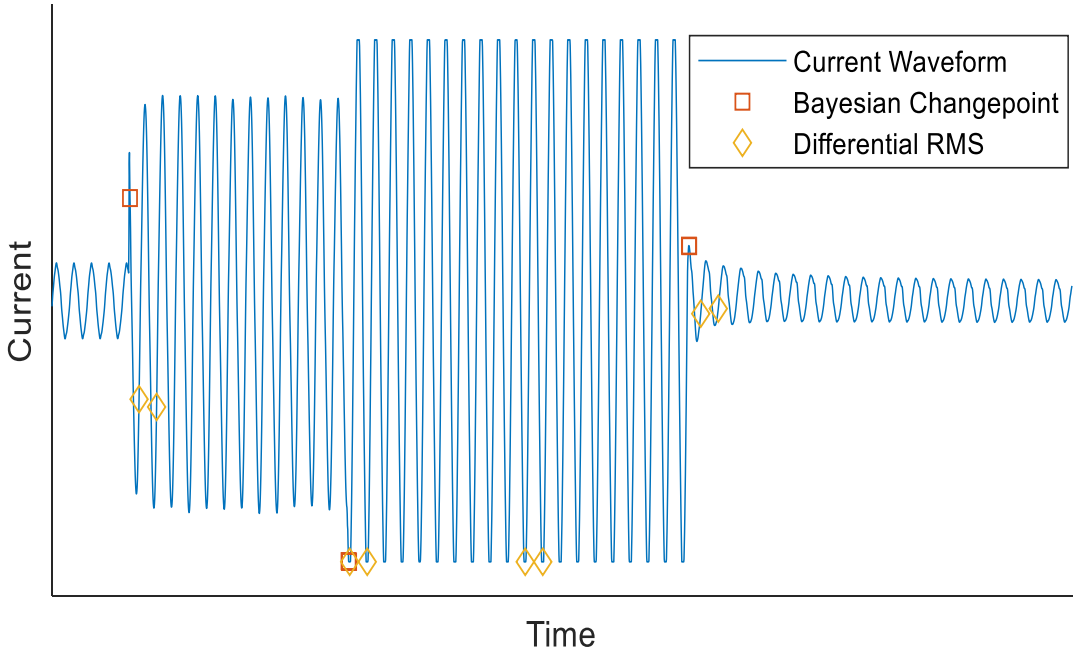



Fig 4-9 Anomaly detection result for BCPQS benchmarked with Differential RMS for case 1 in operational data set

## 4.5 Conclusion

The contribution of this chapter has been to utilise online Bayesian Changepoint methods to automatically detect and segment operational fault and anomaly data collected from distribution networks under typical operation and realistically challenging simulated conditions. Compared with conventional monitoring approaches, the proposed method offers improved false alarm rates, earlier detection of disturbances and also provides associated severity metrics, in terms of observation likelihood, interpretable as fault stage development. In addition, this chapter has introduced a new metric for evaluating the

effectiveness of detecting anomalies online from continuous streams of distribution network monitoring data. Both operational data from a typical distribution network and simulations of various realistic challenges including high power electronic penetrations have been used to demonstrate its effectiveness against state of the art approaches.

The segmentation algorithm, described in this chapter, supports online abnormal waveform extraction by accurately identifying only the periods of interest within continuously monitored data. However, the extracted signal is still difficult to interpret and analyse. Therefore, the next chapter will discuss how to identify the fault signatures using machine learning.

# Chapter 5

# Automated Distribution Network Fault Cause Identification with Advanced Similarity Metrics

## 5.1 Introduction

Chapter 4 has discussed how continuous PQ data streams can be segmented to isolate periods of anomalous operation; however, the fault cause resulting from the selected anomaly waveform is difficult to interpret. As Chapter 3 showed, the ability to identify the causes of disturbances via PQ waveforms is beneficial from both an operational and an asset management perspective. Widespread recognition of the causes of faults over time can allow maintenance for affected assets to be planned and enhance the situational awareness for the network. Therefore, an appropriate fault cause identification approach based on the segmented PQ waveform is required.

Traditionally, fault causes were identified through manual analysis of weather and fault behaviour [16]. The expert knowledge that defines this is difficult to standardize across cause identification and is time-consuming and therefore expensive to undertake. Additionally, the complex form faults can now take, makes this endeavour challenging to define new criteria for fault identification. High-resolution fault and disturbance recording equipment compounds this problem further, in that the waveform level representations they capture are too voluminous to interpret manually.

In response to this, recent works have already moved on to using automatic classifiers: [47] has shown an application of knowledge-based features to accurately identify causes, however, the choice of an appropriate threshold still requires the intervention of a domain expert, which can hamper the scalability of this solution. The works [84][85] respectively proposed Artificial Neural Network (ANN) and Fuzzy Classification using field context data to identify outage causes via training with a large amount of examples. [86] utilised One Nearest Neighbour (1-NN) to rank and validate the relevant contextual and waveform features for transmission level fault identification, while [16] constructed a Deep Belief Network (DBN) to identify fault cause. The automated approaches outlined, such as ANN and DBN, required a large number of exemplars to train the classifier. Most DNOs would

consider archiving curated fault data marked up with diagnostic labels to be beyond their usual remit. Many state of the art classifiers [47][84][85][86] require thousands of examples to learn from which is time consuming and impractical. However, previous research [87][1][70] identified that many faults and failures can exhibit similar characteristics. This would be classed as "event similarity". This provides a possibility to reduce the demand on the labelled examples.

Event similarity could be used to automatically identify a recurring fault situation via patterns learned from this historical data [88], which can in turn be used for diagnosis and prognosis of recurrent incipient faults observed operationally [7][24]. Operational noise and variability make matching events to historical equivalents difficult, necessitating means of similarity to be developed specifically for PQ events.

To support the application of a fault cause identifier for practical use on distribution networks, the following problems need to be addressed: extensive labelled fault exemplars are not always available for training classifiers, therefore, this Chapter proposes a means of inferring fault cause from operational data through analysing the most similar PQ events on a distribution network; fault signatures can vary in duration and magnitude even when they result from the same cause - the proposed approach eschews existing pointwise means of comparison to deal with similar fault cases that may be misaligned; Extraction of relevant features as input to a classifier requires extensive domain knowledge to inform an optimal selection that can accommodate natural variability and context. This is time and resource intensive and even the best feature extraction is still going to discard part of the waveform. The approach proposed here uses all of the data comprising the waveform rather than just a representative feature.

This solution could automatically interpret a segmented disturbance waveform without the need for a large set of exemplars to train classifiers to diagnose faults. The resulting classifier can simply be embedded into the framework in Chapter 3 and propagate the

predicted fault context to maintenance crews who in turn can approach root cause investigations with higher situational awareness. The model capability and performance are demonstrated on the US Department of Energy (DoE) Power Quality data set which has been introduced in Chapter 3. This performance is compared with conventional classifiers drawn from recent literature.

## 5.2 Similarity-based Fault Cause Identification

PQ disturbance causes are multifactorial which presents difficulties in identifying features that represent particular faults [89][86]. The DoE PQ data set provides 166 expert labelled three-phase AC voltage and current signals at 0.96 and 3.84 kHz [46]; one such event is shown in Fig 5-1.

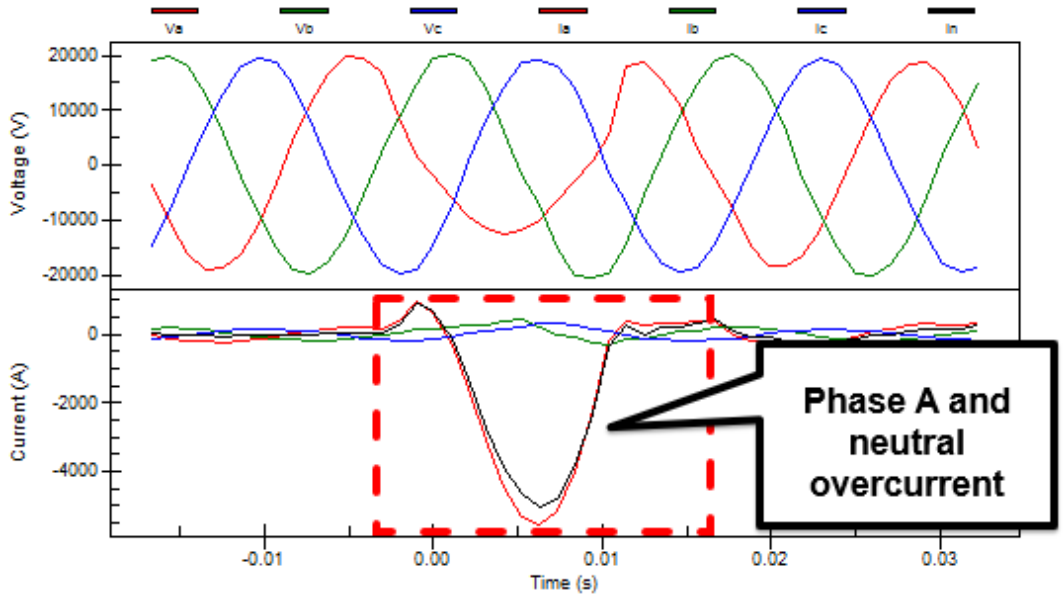



Fig 5-1 Power Quality Waveforms for short term phase-earth overcurrent. The fault clears in 0.042 sec; overhead arrester failure; isolated by recloser; clear weather; happened at 5/19/2005 04:40:26.1990, Phase A

Fig 5-1 shows an example of how a fault may manifest in waveform data. This fault is caused by weather – specifically, heavy snow on overhead lines. During the fault, distortion of one-cycle of voltage on phase A and neutral overcurrent can be observed. This can be

because the snow causes a short-circuit which conducts from the overhead line to ground. An approach of fault causes identification based on the PQ waveform will be discussed below.

The Nearest Neighbour (1-NN) classifier using Euclidean distance as its similarity measurement has been previously validated to classify fault cause with a large amount of training data at transmission level [86]. Here, 1-NN based on a new similarity measurement is proposed to identify the recurrent fault and retrieve associated cause behind the PQ disturbance events but with only a small amount of training data. Fig 5-2 demonstrates the processing stages of the proposed similarity-based classification model.

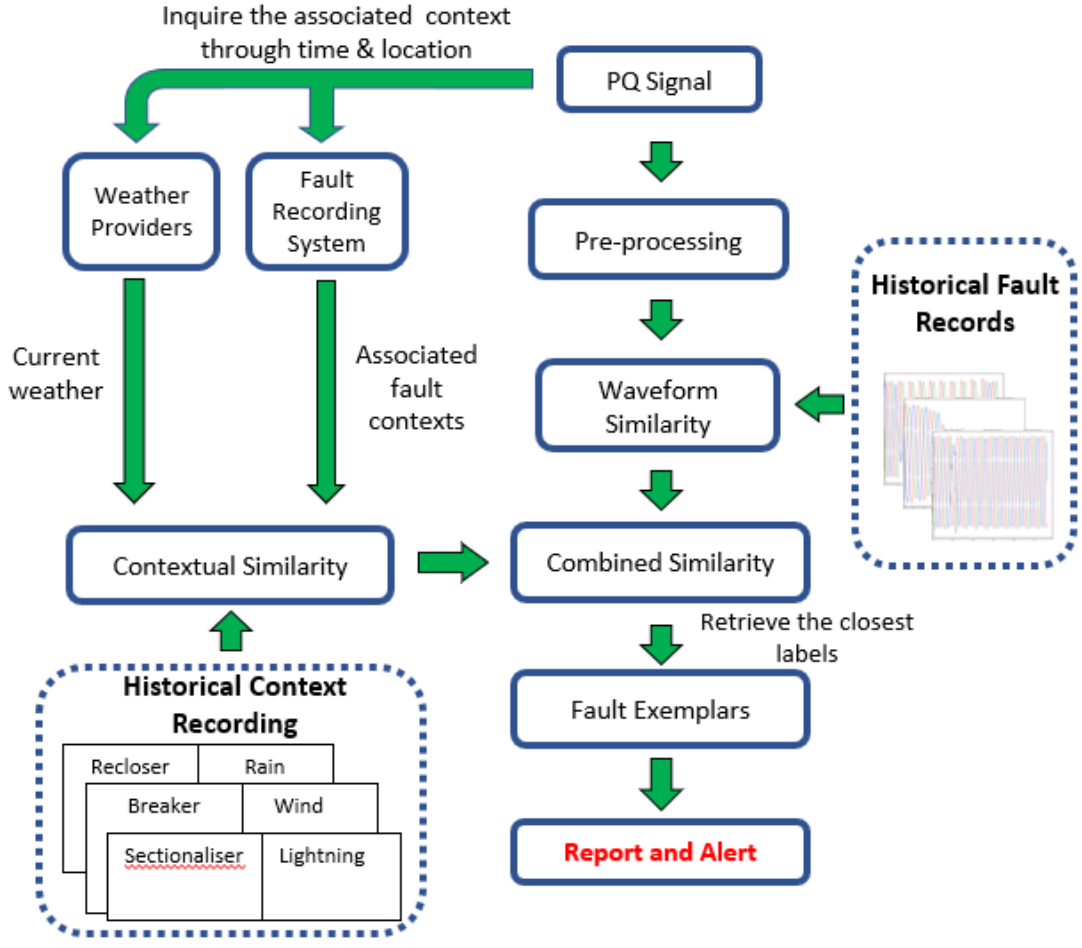



Fig 5-2. Online processing stages for the proposed automatic fault cause identification.

From Fig 5-2, the waveform processing output stages are associated with context obtained from fault recorders, such as isolation equipment operated, and weather providers such as localized environmental conditions. Since faults manifest as abrupt noise signals rather than changes in periodicity, the non-periodic component from the three-phase current is extracted from the raw data through a pre-processing function before evaluating the waveform similarity between event pairs. Beyond this, the similarity of the associated context will be assessed through comparison with the context of historical events. Then a combined similarity measure of the waveform and the context will be inserted into the 1-NN to retrieve the closest historical event and infer the associated fault cause for reporting. The detailed function of these processing stages will now be described.

### *5.2.1 Waveform Pre-processing*

To mitigate the influence of the sinusoidal waveform on the similarity, the fault components can be extracted via removing the sinusoid components. The approach to decoupling the sinusoidal components from the abnormal components of the signal is to superimpose faults over the prior normal cycle waveform [6] which is identical to the method in Chapter 3. Then the shape of the residual fault components can be used to evaluate the similarity. An example of the residual fault component is given in Fig 5-3.

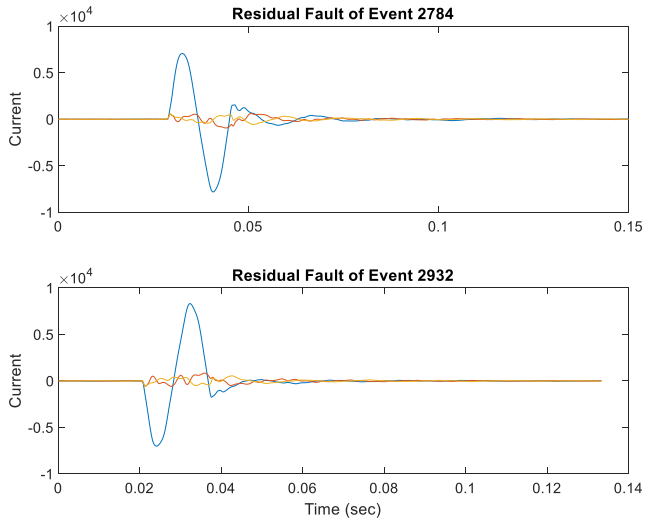



Fig 5-3. Residual fault components of event 2784 and event 2932

As Fig 5-3 shows, some faults, such as arcing, are usually triggered at the peaks [54] and when they initiate at peak or valley positions, this affects the sign of the residual. To solve this, the absolute value of the fault components is used to evaluate the similarity between pairs.

### *5.2.2 Waveform Similarity Measurement*

The duration of instances of the same fault can be different. To eliminate the effect of this, a signal alignment technique is required. Dynamic Time Warping (DTW) is a dynamic programming based time algorithm which has been widely employed to calculate similarity between two signals with different durations [90], such as spoken word, by ignoring both global and local shifts in the time dimension. Assuming two post-processed temporal signals $U$ and $V$ with different duration:

$$U = u_1, u_2 \dots u_n \dots u_N \tag{5.1}$$

$$V = v_1, v_2 \dots v_m \dots v_M \tag{5.2}$$

where $N$ are $M$are the length of the signals and $N \neq M$. To eliminate the effect of different durations, DTW uses a pairwise assessment of amplitudes as the distance between observations in $U$ with observations in $V$. The resulting N by M distance matrix is shown in Fig 5-4 and provides an optimum path from the bottom left to the top right which is called warping path, $WP(k)$, traverses as:

$$\begin{gathered} WP(k) = w(1), w(2) \dots w(k) \dots w(K), \\ \max(N, M) \leq K < N + M \end{gathered} \tag{5.3}$$

where $K$ is the length of warping path, $k$ is the index of the warp function and $w(k)$ is the element of the warp path. To prevent information loss during calculating similarity, a minimized cumulative distances warp path, $D(WP(k))$, is required. The cumulative distances warp path is also called the cost matrix:

$$C_k = \sum_{j=1}^{k} Dist(w(j)) \tag{5.4}$$

$$D(WP(k)) = Min(C_k) \tag{5.5}$$

where $Dist()$ is a distance function, such as Euclidean distance; $C_k$ is the value of the cost function at the $k$th element of the warping path.

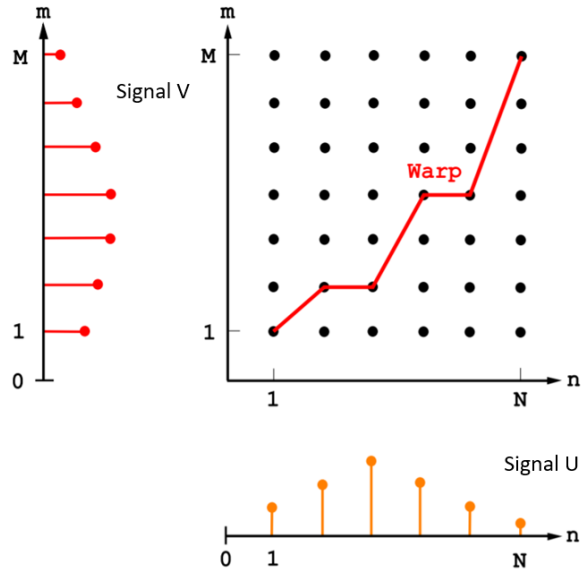



Fig 5-4. DTW cost matrix formation for two signals Y and X of duration M and N; the warping path is defined as the lowest cost route from cell 1,1 to N, M

As Fig 5-4 shows, the warping path starts from (1,1) and ends at $(N, M)$. A constraint requires the warping path to monotonically increase, so the update of the warping path is given as:

$$\begin{aligned} &D\big(WP(k+1)\big) \\ &\quad = D\big(WP(k)\big) \\ &\quad + \min\left(Dist(i, j+1), Dist(i+1, j+1), Dist(i+1, j)\right) \end{aligned} \tag{5.6}$$

The minimized cumulative distance at K, $D\big(WP(K)\big)_{(U,V)}$, represents the waveform similarity between signals $U$ and $V$. DTW will have a higher tolerance to phase distortion

compared to conventional pairwise means of assessing similarity since it carries out the alignment prior to the similarity assessment.

### *5.2.3 Context Similarity Measurement*

With the context of the new fault extracted, contextual similarity can be evaluated. As Fig. 5-2 shows, the context can be extracted through time and location. This paper utilizes the same context data as in [84] [91], which are a timestamp, local weather, isolation equipment and phase affected, as Table I shows. Timestamp can provide season and time of day; interrupting device and fault phase can be provided by SCADA or IED devices; weather data can be provided by a weather service using the specified time and location. All of the proposed contextual data are commonly available. However, context usually takes the form of a label (which can be a categorical value) which makes similarity measures, such as Euclidean distance, unsuitable. To address this, contextual similarity based on the Hamming distance is used as a measure of how closely context is associated with an event. Hamming distance, expressed as $D_H$, has been used to measure the distance between examples that have multiple categories attached to them [92] :

$$D_{H(U,V)} = \frac{1}{N}\sum_{i=1}^{N} |Y_{i(U)} - Y_{i(V)}| \tag{5.7}$$

where $Y_{i(U)}$, $Y_{i(V)}$ are the categories that represent the context of signal $U$ and $V$ respectively. N is the number of contextual features. The output of the Hamming distance is a discrete value.

### *5.2.4 Combined Similarity*

Faults can manifest through the waveform but can also be jointly related to the context they occur in; therefore, combined similarity can be a beneficial approach to indicate the

relations between the fault being investigated and historical events. It is proposed to combine waveform similarity and contextual similarity as follows:

$$Comb_{(U,V)} = \frac{D_{H(U,V)}}{\max\left(D_H{}'\right)} \cdot \frac{D(WP(K))_{(U,V)}}{\max\left(D\left(WP(K)\right)'\right)} \tag{5.8}$$

where $D_H{}'$ is the contextual similarity between historical current events and $D\left(WP(K)\right)'$ is the corresponding waveform similarity.

The weights used for waveform similarity and contextual similarity are identical. However, waveform and contextual similarity contain different amounts of useful information. Therefore, the weights might affect the final result. To demonstrate, weights are added to the combined similarity as:

$$Comb_{(U,V)} = \left(\frac{D_{H(U,V)}}{\max(D_H{}')}\right)^{w} \cdot \left(\frac{D\left(WP(K)\right)_{(U,V)}}{\max\left(D\left(WP(K)\right)'\right)}\right)^{1-w} \tag{5.9}$$

Where $w$ is weight of context similarity which is within [0,1]. $\frac{D_{H(U,V)}}{\max(D_H{}')}$ is the normalized context similarity and $\frac{D(WP(K))_{(U,V)}}{\max\left(D(WP(K))'\right)}$ is normalized waveform similarity. The result of a sensitivity analysis for this is shown below:

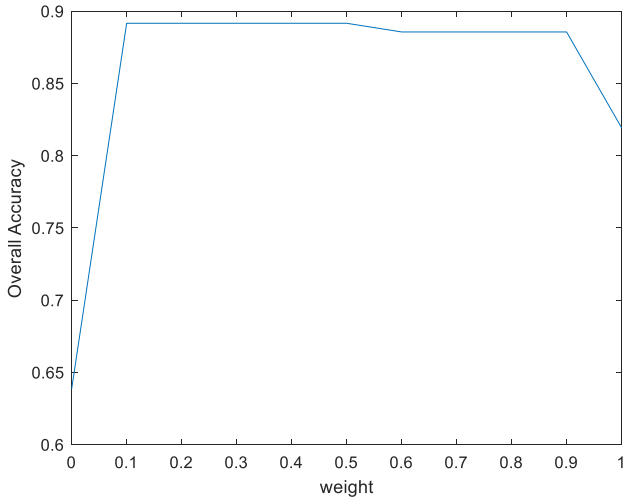


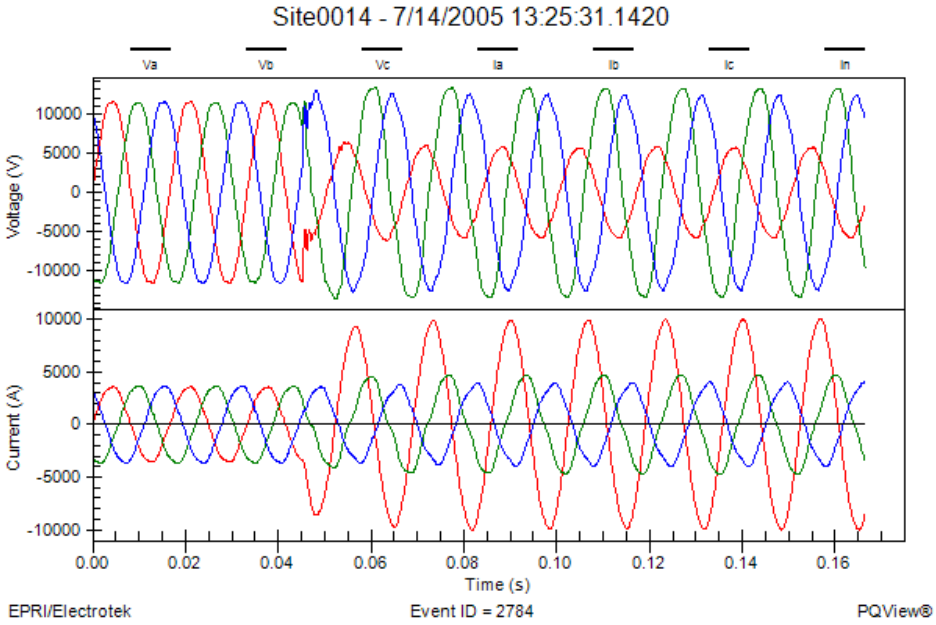

Fig 5-5. Weight analysis for waveform similarity and contextual similarity

As Fig 5-5 shows, the result of using waveform similarity alone is significantly worse than the result of using the context similarity alone. However, higher or identical weight of waveform similarity can result in a slightly higher accuracy than higher weight of context. This can be because the variance of context similarity is higher than waveform similarity, and the waveform similarity carries some very important information. Generally, the weight does not affect the classification result a lot. Therefore, to simplify the similarity, equal weights are used.

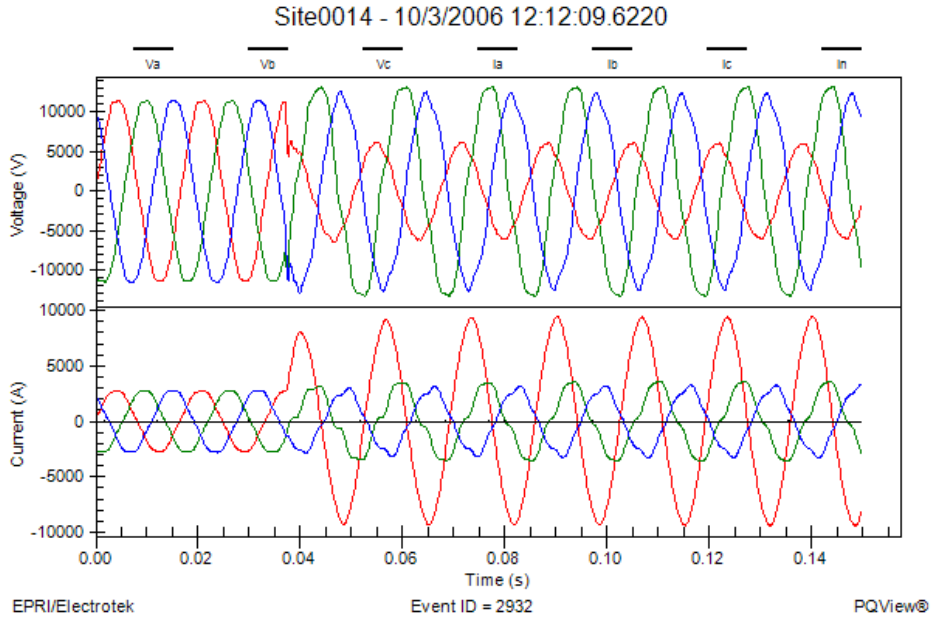
## 5.3 Case Study: Recurrent Fault Identification

In order to validate that the proposed similarity metric can be used to express the relationship between PQ event causes and their waveforms as well as the relationship between PQ event causes and their contextual features, the EPRI DoE Power Quality data set is used [46]. As Chapter 3 described, the data was sourced from various power quality monitors, digital fault recorders, microprocessor relays, and remote terminal units (RTUs). This provides 3-phase voltage and current measurements sampled at 0.96 kHz and 3.84 kHz for 334 power quality fault instances. Among these, 166 faults and disturbance records have been labelled by experts according to their cause, environmental conditions and associated failed plant. Two cases are presented to highlight the practical effectiveness of the metric: the first case validates that the proposed waveform similarity measurement can identify shape-based recurrent faults. The second case is to validate that the proposed contextual similarity can identify recurrent faults based on context. Both cases use the same pair of events, shown in Fig 5-6, for comparison purposes; its residual fault component has been given in Fig 5-3. These two faults in Fig 5-6 (a) and (b) came from the same substation (site 14 in the DoE set), resulted from a terminator failure and the fault was interrupted by a circuit breaker. However, the time interval between these two events in Fig 5-6 is more than one year, the incident report may have been discarded in this time, and numerous subsequent events may have resulted with the cause being forgotten,

preventing domain knowledge from facilitating fault analysis and therefore resolution. Consequently, the PQ waveform of these two events is very similar although it can be seen that event 2932 from Fig 5-6 (a) begins half a cycle earlier than event 2784 from Fig 5-6 (b). It is quite common to see incipient faults, such as arc faults in underground cables [70], begin at different parts of the cycle, which would have rendered a pointwise similarity metric ineffective.




(a) The first terminator failure happened at 14th July 2005




(a) The second terminator failure happened at 3rd Oct 2005

Fig 5-6 two PQ disturbance events with a high similarity. Although the cause is the same in both cases, a pairwise comparison would have overlooked this due to differences in the duration and cycle position of fault initiation [46].

### *5.3.1 Shape-based Signal Similarity*

Using the two examples from Fig 5-6, the signal similarity evaluates the waveform shape differences between two PQ events. Their waveform similarity is evaluated as 0.99 and the maximum value is 1, indicating highly similar events which is in agreement with the actual fault causes.

### *5.3.2 Contextual Similarity*

The fault records used contain associated information which is detailed in Table 5.1. Although the gap between the two faults is more than one year, they occur in similar contexts, such as the time of day and location. The only difference is the season, but the associated ambient temperature on a given day in North America may be the same in Autumn and Winter so this could be uninformative. For this case, the proposed method has calculated the contextual similarity between the two events in Fig 5-6 as 0.8 which is four identical features out of five known features.

Table 5.1 Fault Context Comparison for A Pair of Recurrent Faults

| FAULT | FAULT 1 | FAULT 2 |
|---|---|---|
| SEASON | SUMMER | FALL |
| FAULTED PHASE | PHASE C | PHASE C |
| DAY TIME | 12:25:31 | 11:12:09 |
| INTERRUPTING DEVICE | BREAKER | BREAKER |
| WEATHER | UNKNOWN | UNKNOWN |
| LOCATION | SITE 4, FEEDER 18 | SITE 4, FEEDER 18 |
| CONTEXTUAL SIMILARITY | 0.8 | |

## 5.4 Fault Cause Identification Benchmarks

Using supervised classifiers to automatically identify fault cause [93] still requires domain knowledge to select appropriate input features. Past works used two broad categories of features to identify fault causes in distribution networks: Waveform-based

features [47] and contextual features [84] [91]. The waveform-based features arise from field experience, for example, animal contact is likely to only affect a single phase owing to the native of physical contact. By the same reasoning, a vehicle pole impact can result in multiple phases being affected through the resulting collision of overhead conductors. From these examples, the number of faulted phases can be inferred as a useful indicator of fault cause. These features extracted from domain knowledge on the DoE data in this Chapter are shown in Table 5.2 and have been previously discussed in [47]. Other prior work [84] has incorporated fault context, such as weather, season, faulted phase and time of day to identify the fault cause. Examples of this are listed in Table 5.3. To demonstrate the performance benefits of the proposed classifier, it will now be benchmarked on operational data against the conventional classifiers with both input features.

Table 5.2 Waveform Characteristics Used for Fault Cause Identification [47]

| SYMBOL | EQUATION | DESCRIPTION |
|---|---|---|
| $R_1$ | $\frac{\text{MAX}\ (I_{amax}, I_{bmax}, I_{cmax})}{\text{MEDIAN}\ (I_{amax}, I_{bmax}, I_{cmax})}$ | THE NUMBER OF FAULTED PHASES |
| $R_2$ | $\frac{\text{MEDIAN}\ (I_{amax}, I_{bmax}, I_{cmax})}{\text{MIN}\ (I_{amax}, I_{bmax}, I_{cmax})}$ | THE NUMBER OF FAULTED PHASES |
| $I_f$ | $I_{pkmax} - I_{0pk}$ | FAULT CURRENT COMPONENT |
| $n_f$ | $\sum_{j=1}^{n} \frac{I_{pk(j)}}{I_{0pk}}$ | FAULT DURATION |
| $\alpha_{ATT}$ | $\frac{I_{pkmin}}{I_{0pk}}$ | FAULT CURRENT ATTENUATION |
| E | ENERGY(WAVEDEC(VNORM)) | FREQUENCY DOMAIN ENERGY PERCENTAGE |

$I_{amax}$ $I_{bmax}$ $I_{cmax}$– maximum value of the current phase A, B, C
$I_{pkmax}$- maximum value of peak point of phase current
$I_{0pk}$- normal operation peak current, here use the peak current of first cycle current
$I_{pkmin}$- minimum value of peak point of phase current

Table 5.3 Contextual Features Used for Fault Cause Identification [84]

| Feature | Value |
|---|---|
| Interrupting Device | Recloser, Fuse, Breaker, Sectionlizer, Switch |
| Weather | clear weather, thunderstorm, snow, windy |
| Faulted Phase | A, B, C, BC, AC, AB, ABC |
| Season | spring, summer, fall, winter |
| Day time | day, night |

Day time: 6:00 am – 6:00 pm

Spring: Mar – May; Summer: June – August; Fall: Sep – Nov; Winter: Dec – Feb

## 5.5 Automated Fault Cause Identification

The analysis in Section 5.4 has demonstrated the use of conventional inputs in fault classification, including waveform-based features and contextual features, and explained why these have been selected. However, these features were investigated only with a large number of exemplars and none of the previous works identifies an appropriate classifier that can work using minimal exemplars. To address this, the following section will investigate the predictive power of different fault event features and performance of different fully automated classifiers trained on a relatively small number of fault examples. These will employ the proposed similarity-based classifier benchmarked against the same classifiers with the waveform-based features, contextual features as well as a combination of both. No user tunable parameters are required, which can help to extend this approach to any fault in future, even for low-prevalence faults. The DoE data is labelled according to fault cause, which provides a means of validating the effectiveness of these classifiers. The low prevalence of faults coupled with the small number of exemplars simulates the realistic environment for fault identification. Five categories of faults are considered from a two-year period: Tree (41 examples), Equipment (75), Animal (11), Vehicle (21) and Lightning (17). To test the classifiers with waveform-based and contextual input features, leave-one-out cross validation, which is appropriate for validating small data sets, is used to understand the level of performance that might be expected in operational use. Leave-

one-out cross validation is an extreme approach of the k-fold cross validation. Leave-one-out cross validation uses every event as a test set in turn, and then train the model using the remaining events. In this way, every event can be fully used; therefore, this is appropriate for validating small data sets. Table 5-4 and Table 5-5 show both the combined and individual class accuracies for all classifier tested.

Furthermore, the classifiers tested are ANN[2], DBN[16], Decision Tree, Discriminant, Support Vector Machine (SVM), KNN[86] and Ensemble methods[94][95], these classifiers were either used in past literature or can be applied for minimal data. And the architecture of these classifiers has been optimized, e.g. the ANN used has 4 hidden layers to connected with inputs, and each layer has 500 neurons, which are fully connected by weights. The activation function for each layer is set as an ELU (Exponential Linear Unit) [22]. To achieve a stable accuracy at the end, 100 epochs are pre-set to validate the neural network. The optimisation method is chosen to be ADAM [23] which is a kind of first-order gradient descent optimization method, selected for its fast convergence properties.

Furthermore, two common evaluation metrics, Overall Accuracy (ACC) and F-score are used to evaluate the performance [86]. ACC can indicate the overall performance of the classifiers, but is not adequate for an unbalanced dataset (where the proportions of exemplars are unequal), whereas F-score can reflect the confusion matrix for every class regardless of how prevalent fault cases are. The formula of them are demonstrated as:

$$\text{Overall Accuracy} = \frac{\sum True\ Postives + \sum True\ Negatives}{Total\ Population} \tag{5.10}$$

$$\text{F-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.11}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{5.12}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5.13}$$

### *5.5.1 Classifier Design for Maximum Accuracy*

Prior research [47][84] used waveform-based features and contextual features respectively to identify fault causes in distribution networks. Table 5.4 demonstrates the performance of different classifiers with both features and the combination of the two. Regardless of the classifiers chosen, the rank of the accuracy metrics show that the contextual features perform better than waveform-based features, but worse than the combination of the two. These will now be described.

Table 5.4 Comparison of Choice of Model and Feature Set for Fault Cause Classifier

| Classifier | Feature Set | F-score | | | | | Overall Accuracy |
|---|---|---|---|---|---|---|---|
| | | Tree | Equipment | Animal | Vehicle | Lightning | |
| ANN [84] | Waveform Features [47] | 18.42% | 32.76% | 0% | 8.33% | 19.27 | 19.27% |
| Bagged Tree | Waveform Features [47] | 67.47% | 79.49% | 40% | 66.67% | 52.94% | 69.88% |
| 1-NN [86] | Waveform Features [47] | 67.42% | 66.67% | 20% | 60.47% | 48.48% | 61.44% |
| ANN [84] | Contextual Features [84] [85] | 22.86% | 28.8% | 10.81% | 22.64% | 12.77% | 22.22% |
| Bagged Tree | Contextual Features [84] [85] | 78.57% | 78.67% | 60% | **88.37%** | 75.68% | 78.9% |
| 1-NN [86] | Contextual Features [84] [85] | 65.71% | 82.67% | 76.19% | 86.38% | 63.86% | 76.6% |
| ANN [84] | Combined Features [84] [47] | 22.78% | 39.62% | 21.26% | 21.05% | 14.04% | 24.09% |
| DBN [16] | Combined Features [84] [47] | 0% | 62.24% | 0% | 0% | 0% | 43.43% |
| Bagged Tree | Combined Features [84] [47] | 77.92% | 82.72% | **73.68%** | 85% | 82.35% | 81.32% |
| 1-NN [86] | Combined Features [84] [47] | **80.95%** | **86.9%** | 66.67% | 79.07% | **83.33%** | **82.5%** |

Table 5.5 Comparison of Choice of Similarity Measure for Fault Cause Classifier

| Classifier | Similarity Measure | F-score | | | | | Overall Accuracy |
|---|---|---|---|---|---|---|---|
| | | Tree | Equipment | Animal | Vehicle | Lightning | |
| 1-NN | Waveform-based similarity | 75% | 65.73% | 33.33% | 52.63% | 61.54% | 63.86% |
| 1-NN | Contextual similarity using Hamming distance | 69.44% | 85.14% | **76.19%** | **90.48%** | 65.31% | 78.9% |
| 1-NN | Combined Similarity | **89.16%** | **90.54%** | 75% | 88.38% | **94.12%** | **89.15%** |

#### *5.5.1.1 Waveform Characteristics*

Although 1-NN and Bagged Tree can identify the faults to a reasonable level (> 60%) in Table 5.4, some fault classes with significant waveform variability (e.g. animal and vehicle) obtain a low F-score. Some of the fault events in the DoE data set manifest over several waveform occurrences. Although the root cause is the same, the waveform shapes can vary drastically. An example is demonstrated in Fig 5-7. The events in Fig 5-7 recur consecutively within a short period and they are both caused by animals; both occur in similar contexts but the waveform looks significantly different. However, through observing the whole dataset, the animal related faults in DoE data set all occur around April to August and frequently occur under fair weather, which means the contextual features can be a more powerful predictor than waveform on these faults. Generally, the waveform characteristics alone can be difficult to use to identify fault causes, especially for faults with significant variability.

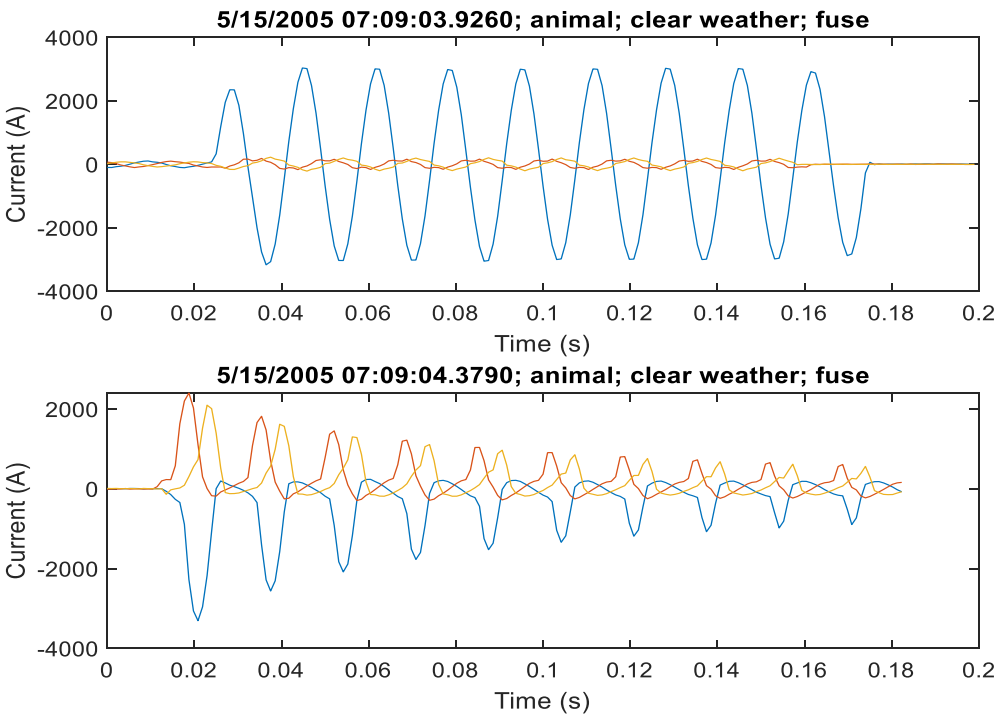



Fig 5-7 Two consecutive animal fault episodes occurring less than a second apart; the waveform is dissimilar but context matches exactly

#### *5.5.1.2 Contextual Features*

Compared to waveform-based classification alone, the use of contextual features alone can improve the overall accuracy by approximately 10% for 1-NN (from 61.44% to 76.66%) and Bagged Tree (from 69.88% to 78.9%) as Table 5.4 shows. Generally, contextual features are considered powerful predictors of fault causes [84], however, the accuracy for contextual features alone is sometimes not enough in practical implementation as Table 5.4 shows.

#### *5.5.1.3 Combined Features*

Table 5.4 shows that the classifiers with the combined waveform derived inputs, augmented with features based on context, generally outperform the classifiers that that only use the individual feature sets alone; e.g. 1-NN, Bagged Tree and ANN achieve 82.5%, 81.32% and 24.09% which improve approximately 6%, 2% and 2% respectively. DBN

cannot work with minimal exemplars which is not taken into account (all the data are categorised into one class). This shows that waveform and context together carry additional information to support accurate classification.

### *5.5.2 Comparison of feature-based and similarity-based method*

To investigate the predictive power of the proposed similarity measures compared with conventional waveform-based features and contextual features, three comparison experiments are carried out. As Table 5.5 demonstrates, if machine learning methods with the domain knowledge listed in Table 5.2 are used to classify the fault causes, a reasonable overall ACC (>60%) is obtained with a 1-NN classifier. Using the proposed similarity measure can enhance classifier accuracy by 2% over an equivalent classifier in Table 5.4 using Euclidean distance on conventional waveform features. Furthermore, as the proposed waveform similarity is directly extracted from the waveform shape, it removes the need to select features for fault identification. Moreover, the contextual similarity-based 1-NN can enhance approximately 2% compared with contextual feature-based methodology. Furthermore, the overall best classification accuracy, 89.15%, comes from using the combined similarity measure, which improves accuracy by approximately 7% over conventional combined features. This also removes the need to select statistical features from the waveform. Every fault class attains an F-score of greater than 75%. Generally, the similarity based method can be advantageous regardless of waveform or contextual inputs used. Fault diagnosis using the combination of both similarity measures still can achieve higher accuracy than using one alone.

### *5.5.3 Classifier Performance with Minimal Available Data*

Instead of waiting to collect a large number of exemplars, utilities may prefer an intelligent classifier that works with minimal available data in order to gain value from monitoring as quickly as possible. Suggested in Section 5.5, and as Table 5.4 and Table 5.5 show, regardless of the features chosen, Neural Networks, including ANN and DBN, have

difficulty in classifying fault cause with the minimal data available, with the test result being worse than a random prediction, such as 24.09% overall accuracy for ANN and 43.43% overall accuracy for DBN using combined features. Having optimized the network to consist of 4 layer and 500 neurons for each layer, performance is still low. Among the other classifiers, Bagged Tree and 1-NN achieve the best results using conventional waveform and contextual features, with the highest accuracy achieved around 82% and 89% respectively. Among the five fault classes, lightning related faults achieved the best F-score even though the exemplar support is not the highest. Therefore, 1-NN using the proposed similarity can provide a reliable fault cause classification without manual intervention, knowledge of which can be used to expedite failure rectification.

### *5.5.4 Performance Impact of Sampling Frequency*

In practice, multiple sampling frequencies are used to record PQ events, including 0.96 kHz and 3.84 kHz [46][67]. For conventional fault classification, the sampling frequency can affect the waveform feature extraction then further affect the classification result. As Table 5.6 shows, the events recorded at higher sampling frequency can achieve 94.44% overall classification accuracy with the proposed similarity using 1-NN – an almost 20% gain which justifies the higher resolution of the data.

Table 5.6 Comparison of Waveform Sampling Frequencies

| Sampling Frequency | Overall Accuracy |
|---|---|
| 960 Hz | 77.58% |
| 3840 Hz | **94.44%** |

## 5.6 Conclusion

This chapter has contributed a new similarity measure to identify recurrent faults and created a classification method to automatically identify fault causes associated with Power

Quality events. The highest accuracy achieved with the new approach is 89.15%, using a combination of waveform and contextual similarity and retaining high accuracy even for low-prevalence events. Possible improvements to enhance the accuracy of the classification may be in developing an automatic exemplar generation model to increase historical events for every eventuality, which allows power quality related faults to be identified during the operational implementation without waiting to accrue a significant archive of labelled examples.

# Chapter 6

# Automated Fault Labelling Using Semantic Analysis of Maintenance Tickets

## 6.1 Introduction

The previous chapters have demonstrated an approach to automatically identifying fault causes in a distribution network, which follows other industries' pursuit of leveraging data to enhance operational understanding. The barrier to this is that in order to classify such faults, a set of labelled faults are required for training in the first instance. Explicit labelling is time consuming and requires expertise to identify and articulate fault taxonomies, which can be difficult when the industry records faults manually and infrequently – many automatic fault cause identification related projects are conducted without extensive labelled data. Although a possible approach to the problem described in last chapter is to use a similarity-based method which can work for minimal exemplars, more exemplars can indeed help the classification further enhance the accuracy. Therefore, this chapter provides a solution to automatically extract labels from other sources which will increase the amount of exemplars for training.

## 6.2 Maintenance Tickets

Ticket based maintenance records and directives exists in a number of service and infrastructure industries. In distribution network operation, often attached to faults are incident or maintenance tickets submitted for validation or work scheduling purposes. These too are typically free text, with a description provided by the individual who filed them and as such will not contain standardized terms or descriptions; instead, it will contain the perspective of the filing individual making it susceptible to ambiguity, personal perspective and hence unusable for supervised machine learning of fault diagnoses; however, these free text usually contain numerous terms relating to useful expert knowledge, such as procedure of past failures or outages, the inspection of each piece of equipment or even transcribed phone conversations with customers. One well-curated example of such a set of incidents, is the EPRI/DoE National Database of Power System Faults [2], 13 examples of which are given in Table 6.1.

Table 6.1 Maintenance records, fault cause labels and associated weather

| EventId | Cause | Weather | Details (free text) |
|---|---|---|---|
| 0001 | Tree | Clear Weather | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 0004 | Tree | Clear Weather | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 0005 | Tree | Clear Weather | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 0007 | Tree | Clear Weather | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) |
| 3042 | Equipment | Unknown | Equipment, Device UG, Damaged. |
| 0021 | Equipment | Clear Weather | Overhead Insulator Failure. BROKEN INSULATOR |
| 0022 | Equipment | Clear Weather | Overhead Insulator Failure. BROKEN INSULATOR |
| 0062 | Undetermined | Raining | Storm |
| 0064 | Undetermined | Raining | Storm |
| 0067 | Tree | Thunderstorm | Tree/Limb Growth |
| 0065 | Tree | Thunderstorm | Tree/Limb Growth |
| 0068 | Tree | Clear Weather | VINES ON TRANSFORMER |
| 2760 | Unknown | Unknown | Short duration variation. No outage information found. |
| 3048 | Equipment | Unknown | Equipment, Capacitor Station, Damaged. |

This data set is unique in that it provides both the maintenance report ('Details') as free text as well as a ground truth classification ('Cause'); operationally, providing these classifications would be an unfeasible effort, so a means of automatically inferring these from the routinely available maintenance notes would be valuable, and this data set provides the means of validating such a method.

## 6.3 Documentation Topic Models

In machine learning, a topic model is a statistic model which is designed to extract the abstract topic from a collection of documents. Topic models have previously been used to understand the underlying problems from unstructured ticket text in distribution networks: [96] leverages content in maintenance records of Manhattan electric networks to rank the most vulnerable areas. [97] used Hidden Markov Model based models to analyse operational free-text data to aid fault identification. This chapter demonstrates an application of leveraging free-text database in power networks to automatically label PQ events by extracting semantic information from the free-text maintenance tickets. Without a topic based model to extract sematic information, the labelling of fault occurrences using selected keywords from maintenance tickets would be prone to spelling, grammatical, style and terminology aberrations which could only be overcome by enforcing strict maintenance reporting guidelines which provides an additional burden on the field operative. A representation popular in the Natural Language Processing and Information Retrieval communities for many years, the 'bag of words' is highly suited to incident tickets and operative fault reports [96]: this entails 'stopping' the document (ticket) by removing common words, stemming all verbs and adverbs (which turns them into a corresponding noun) and leaves the document as a vector of word occurrence counts. This approach yielded a number of widely used document similarity metrics based on distances between these vectors that reflected commonality of terms. Subsequent probabilistic formulations of this approach could be used to imply polysemy and synonymy among terms making them ideal for identifying documents with the same sentiment but different term usage [98] – a characterizing problem of maintenance reports.

Latent Dirichlet Allocation (LDA) [99] is a probabilistic model which can extract semantic representations using a hypothetical topic distribution. The model assumes each document can be represented by a different mixture of latent topics, and each topic can be

characterized by a distribution over words. To achieve this, LDA model utilises the following probability distributions:

- Dirichlet distribution based on hyperparameter $\alpha$ is used to generate a mixture topic distribution $\theta$, which is multinomial distributed
- Multinomial distribution $\theta$ is used to select the topic $z$ of each of $M$ documents
- Dirichlet distribution based on hyperparameter $\beta$ is used to determine distribution over $N$ words for the selected topic $z$, which generates the probability for word $w$

The diagram is demonstrated as:

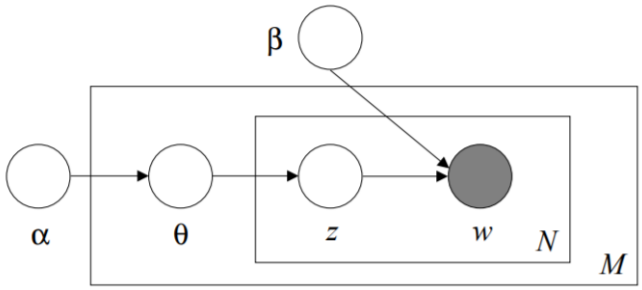



Fig 6-1 Plate notation for LDA model; dark circle represents observations

The equation can be expressed as:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\theta_d) p(w_{dn}|Z_{dn},\beta) d\theta_d \qquad (6.1)$$

Where $\alpha$ and $\beta$ are the hyper-parameters of two Dirichlet distribution; $d$ and $M$ represent the current index and total number of documents, $n$ and $N$ represent the current index and total length of words in a document. After an appropriate optimization, such as EM algorithm [99] and Gibbs Sampling [100], the topic distribution can be obtained.

The use of a Dirichlet distribution with parameter α, selects the proportion or composition θ of topics in a given document:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{M}\alpha_i)}{\prod_{i=1}^{M}\Gamma(\alpha_i)}\prod_{j=1}^{M}\theta_j^{\alpha_j} \tag{6.2}$$

While β similarly parameterizes a conditional Dirichlet distribution of words over each topic.

## 6.4 The Relation Between Expert Labels and Maintenance Ticket Content

Using LDA to process maintenance reports, a representation of the topic distribution can be obtained. Then a small selection of state of the art classifiers were trained with to validate the relations between expert labels and maintenance tickets using DoE data [50]. The detail is demonstrated in Fig 6-2.

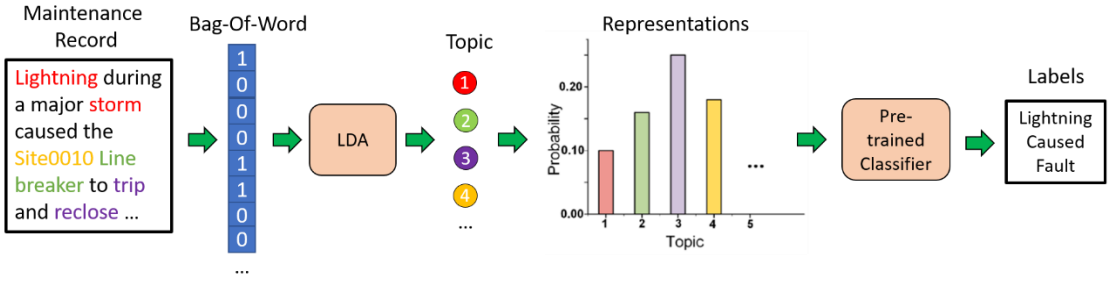



Fig 6-2 an example of using LDA to extract labels from document using supervised learning

Each maintenance record is converted to the bag of words representation and then the resulting word vector is run through the LDA model to get a topic vector associated with each fault record and feed them in a pre-trained classifier produce its label. For predicting the expert label from just a topic vector, Table 6.2 shows the accuracy (the ratio of true positives plus true negatives to all classifications made) of 10 classification models, all of which work on different discriminatory principles and decision surface shapes.

Additionally, the number of latent topic variable is selected as five, which could be interpreted as fault causes types.

Table 6.2 Classification Performance of 5-Topic LDA Model

| Classifier | Maintenance Ticket Label Prediction Accuracy | Parameters |
|---|---|---|
| Ada Boosted Tree | 54.7% | Learning rate =1, The maximum number of estimators=50 |
| Decision Tree | 76.2% | The minimum number of samples required to split an internal node=2, The minimum number of samples required to be at a leaf node=1 |
| Gaussian Process | 61.9% | Alpha=1e-10 |
| Linear SVM | 59.5% | penalty ='l2', loss= 'squared_hinge', tolerance=1e-4, Regularization parameter =1 |
| Naive Bayes | 45.2% | variance_smoothing = 1e-9 |
| Nearest Neighbour | 78.6% | Number of neighbors=5, Weights= 'uniform', Leaf size= 30, Metric='minkowski' |
| Neural Network | 69% | Layer =1, hidden_layer_sizes = 100, activation='relu', solver='adam', alpha = 0.001, learning rate=0.001, Maximum number of iterations=200, Tolerance=1e-4, Momentum=0.9, |
| QDA | 45.2% | Regularizes the covariance=0 |
| RBF SVM | 66.6% | Regularization parameter=1, Kenel='rbf', Degree of the polynomial kernel function=3, Gamma=scale, Tolerance=1e-3, |
| Random Forest | 73.8% | The number of trees in the forest=100, The minimum number of samples required to split an internal node=2, The minimum number of samples required to be at a leaf node=1, The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node=0 |

Using a 25% held out set from a selection of 168 labelled examples, Table 6.2 shows that given an appropriate classifier choice, the topic composition vector provided by the LDA model can be related, and is therefore implicit of the sentiment conveyed in the maintenance report since it corroborates with the label provided by the domain expert in the DoE data set. The relation has been proved; however, the supervised methods results in another demand on labels for maintenance ticket which is not acceptable. Therefore, the next logic step is to further reduce the burden of collecting labelled maintenance tickets by replacing the supervised classifiers with an unsupervised classifier which does not require labelled documents.

## 6.5 Exemplar Generation Using Fault Incident Ticket

As Chapter 3 described previously, the associated labels can be extracted from maintenance tickets during offline implementation. To reduce the burden of pre-processing maintenance tickets data, an unsupervised classifier is required. The fault labels and the associated maintenance tickets are free-text in nature; therefore, a semantic meaning extracted from both documents should be identical; however, the order of words is not taken into account because of Bag-of-Word. This provides a possibility to utilise two documentation topic models to extract both semantic meaning from maintenance records and expected labels, then a document similarity measure can be used to match the most semantically similar maintenance records and labels. Details of this are given in Fig 6-3.

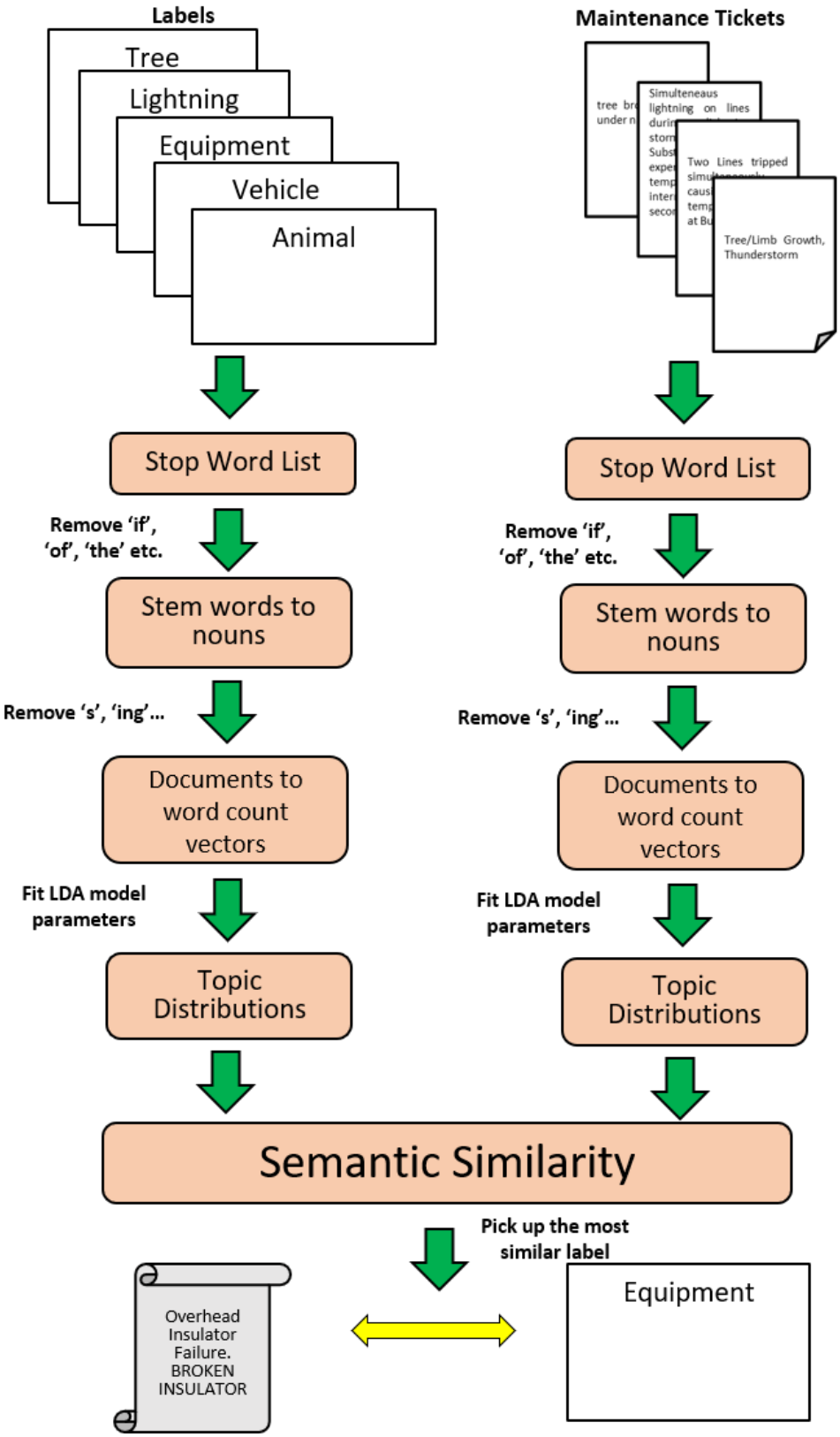



Fig. 6-3. Process for automatically labelling faults with maintenance records

As Fig 6-3 shows, the stop words are removed and remaining words are stemmed, and the topic distributions are generated for the labels and maintenance reports, then the ticket similarities are calculated using semantic similarity. Fig 6-3 shows how the topic distribution would be generated and then associated with fault records – the resulting fault would be automatically labelled with the most similar given topic. As with the data pre-processing used in [50], both labels and maintenance tickets are converted into topic distributions by using the same bag of words and LDA model; then the maintenance tickets will be associated through calculating their semantic similarity in Vector Space Model [101]. The maintenance tickets and labels can be expressed as the vectors:

$$L = l_1, l_2, l_3 \dots l_j \dots l_J \tag{6.3}$$

$$M = m_1, m_2, m_3 \dots m_h \dots m_H \tag{6.4}$$

Where $l_j$ and $m_h$ denotes the $j$th label and $h$th maintenance ticket. Each label and maintenance ticket also can be decomposed into topic distribution through extracting semantic features by LDA model, which can be expressed as:

$$l_j = f_1, f_2, f_3 \dots f_n \tag{6.5}$$

$$m_h = f_1', f_2', f_3' \dots f_n' \tag{6.6}$$

The relations between each pair can be evaluated by using the cosine similarity as demonstrated below:

$$Similarity = \cos(l_j, m_h) = \frac{\sum_{n=1}^{N} f_n f_n'}{\sqrt{\sum_{n=1}^{N} f_n^{\,2}} \sqrt{\sum_{n=1}^{N} f_n'^{\,2}}} \tag{6.7}$$

Where $f_n$ and $f_n'$ are the $n$th topic probability for $l_j$ and $m_h$.

In this case, each event has five possible fault causes. The most similar label will be attached to the maintenance ticket. Additionally, since both the maintenance tickets and fault records, such as historical fault waveform and their context, can usually be associated

by using the timestamp and the location for the fault. Therefore, the labelled fault records data can be synthesized with these relations as Fig 6-4 demonstrates.

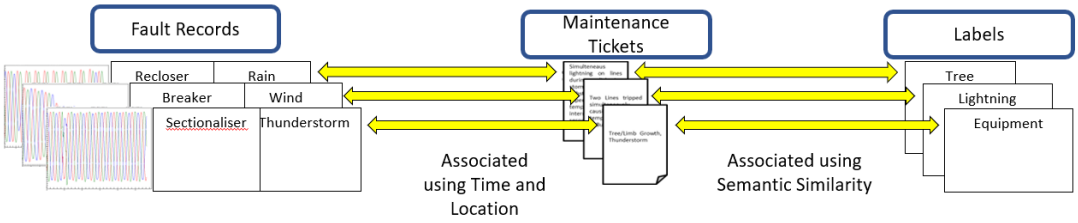



Fig 6-4 the relation between maintenance tickets, fault records and domain expert assigned labels

## 6.6 Automated Labelling Performance

The key barrier to applying supervised machine learning techniques for fault diagnosis in power systems applications is the effort required to produce a sets of labelled exemplars for models to learn from. This section demonstrates how a topic model implemented with realistic data to extract labels for the faults. The first 14 examples of similarity calculation are demonstrated in Table 6.3: 1 represents perfectly relevant and 0 means completely irrelevant. Most events are strongly relevant to one topic. However, events 3024, 0021, 0022 and 3048 are attributed to both Vehicle Impact Fault (VIF) and Equipment Caused Fault (ECF), and more strongly associated to ECF. This is because many VIF events in the operational dataset usually damages the poles and overhead line. The most similar label for faults is used to train the cause classifiers. Since maintenance tickets are produced as a consequence of routine operational procedure, this removes the bottleneck associated with translating domain knowledge into machine learned profiles without the need for manual labelling.

Table 6.3 The Semantic Similarity Between Maintenance Reports and Labels for First 14 Cases, T – Tree, A – Animal, L – Lightning, V – Vehicle, E – Equipment

| Event id | Details (free text) | T | A | L | V | E |
|---|---|---|---|---|---|---|
| 0001 | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) | 0.97 | 0.15 | 0.20 | 0.09 | 0.08 |
| 0004 | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) | 0.97 | 0.15 | 0.20 | 0.09 | 0.08 |
| 0005 | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) | 0.97 | 0.15 | 0.20 | 0.09 | 0.08 |
| 0007 | Fault caused line recloser lockout. Tree Outside Right of Way (Fall/Lean On Primary) | 0.97 | 0.15 | 0.20 | 0.09 | 0.08 |
| 3042 | Equipment, Device UG, Damaged. | 0.22 | 0.15 | 0.22 | 0.93 | 0.94 |
| 0021 | Overhead Insulator Failure. BROKEN INSULATOR | 0.21 | 0.14 | 0.21 | 0.93 | 0.94 |
| 0022 | Overhead Insulator Failure. BROKEN INSULATOR | 0.21 | 0.14 | 0.21 | 0.93 | 0.94 |
| 0062 | Storm | 0.37 | 0.94 | 1 | 0.56 | 0.27 |
| 0064 | Storm | 0.37 | 0.94 | 1 | 0.56 | 0.27 |
| 0067 | Tree/Limb Growth | 0.98 | 0.17 | 0.25 | 0.15 | 0.12 |
| 0065 | Tree/Limb Growth | 0.98 | 0.17 | 0.25 | 0.15 | 0.12 |
| 0068 | VINES ON TRANSFORMER | 0.81 | 0.27 | 0.37 | 0.24 | 0.38 |
| 2760 | Short duration variation. No outage information found. | 0.20 | 0.89 | 0.97 | 0.44 | 0.08 |
| 3048 | Equipment, Capacitor Station, Damaged. | 0.28 | 0.18 | 0.28 | 0.90 | 0.98 |

To demonstrates the effectiveness of the exemplar generation model, the confusion matrix from a classifier trained using the topic generated labels and actual labels is given in Table 6.4.

Table 6.4 Confusion Matrix of Fault Labelling Using Document Similarity of LDA Model (74.69% Overall Accuracy)

| Actual / Predicted | Tree | Equipment | Vehicle | Animal | Lightning |
|---|---|---|---|---|---|
| **Tree** | 41 | 0 | 0 | 0 | 0 |
| **Equipment** | 7 | 51 | 7 | 8 | 2 |
| **Vehicle** | 0 | 0 | 21 | 0 | 0 |
| **Animal** | 0 | 3 | 0 | 6 | 0 |
| **Lightning** | 11 | 0 | 2 | 0 | 5 |

As Table 6.4 shows, approximately 74% labels agree with actual labels (true positive/ total instances). Among these fault classes, Tree Contact and Vehicle Impact Fault can be classified to a high degree of accuracy (Recall = 100%). Equipment Cause Fault and Lightning Strike Fault have many false classifications. Among them, the Lightning Related Faults are even worse than random which can be because some Tree Contact Events are related to the lightning strikes which confuse the LDA model, two examples are demonstrated as Table 6.5 shows.

Table 6.5 Instances of Tree Contact Related With Lightning Related Fault

| Maintenance Record | Fault Cause Label |
|---|---|
| Transmission Line tripped during a major storm. The cause of this event was likely tree contact. Breakers at Substations tripped and reclosed multiple times. | Tree |
| Lightning struck transf, burned primary down.Overhead Primary Failure (No Tree Involved) | Lightning |

Both instances records contain key word 'Tree'. The second record intended to address no tree involved; however, this can confuse the topic model to identify the fault cause – only nouns are taken account for document similarity calculation.

Two solutions to this problem will be investigated in future: 1) a documentation topic model with a stronger robustness to noise is required, and 2) more maintenance records of unrelated tree might be required for LDA model to obtain the representations.

Generally, compared to the result of using supervised learning in Table 6.2, the accuracy is comparable but the method contributed here does not require labelled documents to train classifiers.

## 6.7 Conclusion

This chapter has proposed a means of automatically labelling historical power system faults by modelling the semantic content in their associated maintenance tickets; this would deal with the bottleneck associated with producing training data for supervised learning of fault classifiers – with a readable description associated with a fault records, there would be no need for engineers to manually label exemplars. Performance of around 75% for predicting fault cause from inferred document semantic content, which suggests that LDA models need larger corpora to learn from: LDA as originally formulated does not lend itself to learning word distributions from short documents i.e. maintenance tickets. Rather than imposing verbosity limits [10] and language guidelines on the filing of maintenance

reports, an LDA model instead may be pre-trained on semantically related documents such as maintenance manuals or abstracts.

# Chapter 7

# Validation of Integrated High Frequency Fault Diagnosis

## 7.1 Introduction

The previous chapters discussed how the anomaly segmentation, fault cause identification and fault label generation methods were designed and constructed; however, an intelligent fault analysis requires concurrent operation of these three functions and every function can affect the final classification result. This chapter will investigate the performance of the whole analysis end-to-end including signal segmentation, fault recognition and fault labelling. This is illustrated using the DoE PQ Events Repository data.

## 7.2 Performance of the Integrated Fault Diagnosis Approach

This section uses DoE operational data to validate anomaly segmentation and labelling functionality, then tests classification with automated labelling which is benchmarked against the classification with expert labelled data. To achieve this, 166 labelled faults are used to go through the system from end to end to test the capability of the whole system. The performance of both models will be measured using the same overall accuracy metric as used in Chapter 5.

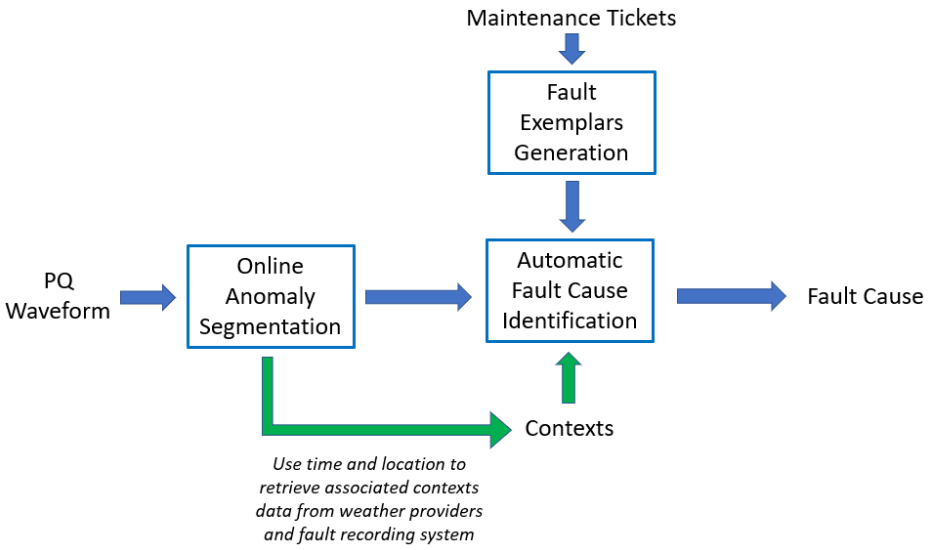



Fig 7-1 The integrated process of the proposed continuous high-resolution fault diagnosis method

The Fig 7-1 demonstrates the high-level of the relations among the models. 166 labelled faults are used to go through the integrated system from end to end to test the capability. Firstly, the faults need to be filtered out by signal segmentation. Then the topic model described in Chapter 6 is utilised to generate a repository of historical exemplars. Topic models usually capture the proportions of topics present in a document and assign the most similar label to the fault. As Fig 3-9 (b) shows, the historical maintenance reports are linked with pertinent fault waveform records using timestamp. And the relevance between every maintenance report in training set with the defined labels, such as 'tree' (this represents tree contact faults), can be calculated using a pre-trained topic model. The most strongly relevant will be selected as the label of the fault. Ultimately, the training set with generated labels will be put into the fault cause classifier, which is 1-NN as Chapter 5 shows, to calculate the performance and benchmark it against the ground truth.

Table 7.1 Confusion Matrix of Anomaly Detection Model (100% Overall Accuracy)

| **Actual Fault** | **Abnormal** | **Normal** |
|---|---|---|
| **Abnormal** | 166 | 0 |
| **Normal** | 0 | 0 |

Table 7.1 shows the performance of anomaly detection using DoE archived fault data, none of the signals are missed by the proposed detector. This validates the effectiveness of the online anomaly detection model again. After segmentation, the remaining data are used to do fault diagnosis. Two situations are compared - the confusion matrix of the classification using ground truth and generated labels on the DoE data for training are shown in Table 7.2 and Table 7.3 respectively. Although trained on faults labelled by different means (automated and manual), both classifiers are tested against the actual labels provided by an expert.

Table 7.2 Confusion Matrix of Fault Diagnosis With Expert Labels (89.16% Overall Accuracy)

| Actual Fault | Tree | Equipment | Vehicle | Animal | Lightning |
|---|---|---|---|---|---|
| **Tree** | 37 | 4 | 0 | 0 | 0 |
| **Equipment** | 3 | 67 | 2 | 3 | 0 |
| **Vehicle** | 0 | 1 | 19 | 1 | 0 |
| **Animal** | 0 | 1 | 1 | 9 | 0 |
| **Lightning** | 2 | 0 | 0 | 0 | 16 |

Table 7.3 Confusion Matrix of Fault Diagnosis With Automatically Generated Labels (71.08% Overall Accuracy)

| Actual Fault | Tree | Equipment | Vehicle | Animal | Lightning |
|---|---|---|---|---|---|
| **Tree** | 37 | 3 | 0 | 0 | 1 |
| **Equipment** | 5 | 52 | 8 | 9 | 1 |
| **Vehicle** | 1 | 0 | 19 | 0 | 1 |
| **Animal** | 0 | 2 | 2 | 6 | 1 |
| **Lightning** | 12 | 0 | 2 | 0 | 4 |

As Table 7.2 and Table 7.3 shows, working with leave-one-out cross validation, the overall classification accuracy against ground truth labels is around 89.16%, and the overall classification accuracy of generated labels is 71.08% - this is an approximately 18% reduction in classification accuracy to achieve fully automated fault labelling using 1-NN with the US Department of Energy data. Furthermore, the reduction stems from Equipment Caused Fault (ECF) and Lightning Strike Fault (LSF), which is consistent with Table 6.4 in Chapter 6 – the reduction results from the labelling error. Therefore, future work should focus improving the fault labelling accuracy.

## 7.3 Conclusion

This Chapter integrates anomaly segmentation, fault cause diagnosis and fault labelling to identify the fault cause from continuous fault waveforms. Archived operational faults are used to test the approach end-to-end. Assuming all archived fault causes labels are the ground truth, the integrated system can achieve 89.16% overall accuracy with minimum volumes of exemplar faults. If a DNO owns numerous maintenance reports rather than exemplar faults, the DNO can still automatically identify faults. However, the accuracy will be reduced by 18% compared to using ground truths.

# Chapter 8

# Conclusions and Further Work

## 8.1 Summary of Contributions

DNOs are experiencing numerous faults on distribution networks, which result in Customer Interruptions (CIs) and Customer Minutes Lost (CMLs). However, due to the extremely low observability of distribution networks, it is challenging for DNOs to make an effective response in timely manner. To increase levels of situational awareness in distribution networks, numerous fault analysis methods are used. Conventional fault diagnosis used fault data from existing devices, such as SCADA and IEDs. However, these methods can only provide limited fault information to DNOs. This is because these devices usually obtain data at regular intervals rather than on a continuous basis, however, numerous incipient faults are intermittent and transient in nature – they can be easily missed. Therefore, there are motives to move to an effective continuous high-frequency fault diagnosis approach. The state-of-the-art research utilised a knowledge-driven approach to continuously identify faults. However, the knowledge-driven approach needs a lot of time to manually capture knowledge for different faults, especially for low-prevalence faults. Additionally, only a limited range of faults can be identified because many faults are difficult to characterize. Therefore, a data-driven approach is needed to solve these limits. However, a continuous high-frequency data-driven approach contains new data related challenges for fault diagnosis: continuous fed-in data will contain large volumes of useless information pertaining to normal operation, which can result in needless storage; data-driven approaches need a large amount of data to pre-train on before identifying fault causes. To address these challenges, this thesis investigated an innovative continuous high-resolution fault diagnosis approach [102] which employs an online segmentation model to extract abnormal waveforms from streamed current data, then the extracted waveform and its associated context are utilised to identify the fault cause.

Contrary to conventional methodologies using local statistical features to detect anomalies, the proposed online segmentation model [103] identifies the start and end

points of faults through tracking residual fault components. The new segmentation model is more robust to the non-linear noise and the false detection is lower. Additionally, to measure the time errors of the detection, a novel evaluation metric is used.

To reduce the demand for labelled data, this research proposed two approaches: the first approach is to use the best classifier which can run with minimal amounts of exemplar faults to identify the fault cause [104]; the second approach is to use an additional model to generate more exemplar faults with a common third-party data source (maintenance reports) [50], then the fault diagnosis can be pre-trained by generated labelled data to identify fault causes. Through integrating a fault segmentation model with a fault diagnosis model, the former method can achieve approximately 89% accuracy on fault cause classification; With the fault labelling model is integrated in, the latter method can achieve roughly 71% on the same data. Although the latter method has a significant reduction on the accuracy, it eliminates the requirement for manual fault labelling - this makes the method scalable and therefore operationally viable.

## 8.2 Future Work

#### *8.2.1.1 Performance Improvement of The Fault Diagnostic Method*

Although the similarity-based fault diagnostic method discussed in Chapter 6 can achieve approximately 89% overall accuracy, higher accuracy would be expected in practical implementation. To achieve this, one of the approaches is to add more contextual information in similarity calculation, which then can provide more discriminative power for the classifier to identify fault causes. Therefore, the next steps can be:

- To gather more field data to investigate whether other features, such as network configuration, can be used to improve the performance.
- To investigate new data generation method, such as time stretching and Generative Adversarial Network (GAN), to increase the diagnosis accuracy.

#### *8.2.1.2 Accuracy Improvement of The Integrated Approach*

As Chapter 7 and 8 discussed, the accuracy of the fault labelling function now is still limited, which then affects the final performance of the end-to-end system. Therefore, one of the areas of future work is to enhance the capability of the fault labelling function. If all generated labels are correct, which can potentially increase accuracy by approximately 18%. A few potential solutions to improve the word representation vector and LDA model can be investigated:

- As Chapter 6 discussed, Bag-of-Words does not take account the order of words and negative words (e.g. not). Therefore, the next step can be to find a new word representation vector which contains this information.
- LDA as originally formulated does not lend itself to learning word distributions from short documents i.e. maintenance tickets. Rather than imposing verbosity limits [10] and language guidelines on the filing of maintenance reports, an LDA model instead may be pre-trained on semantically related documents such as maintenance manuals or abstracts.
- An alternative NLP technique can be used such as Probabilistic Latent Semantic Analysis (PLSA).

#### *8.2.1.3 Hardware Based Implementation*

As the concept of continuous high-resolution based fault diagnosis and the performance of each function has been validated, the next logical step should be using an appropriate hardware platform to validate the effectiveness. This requires extensive practical knowledge, such as data acquisition and parallel computation, to set up an experiment. The algorithms, including anomaly segmentation model, fault cause identification model and fault labelling model, can be adjusted during the hardware testing.

After tuning models for online implementation, the whole work can move to the next Technology Readiness Level (TRL) which is to produce a prototype to demonstrate the performance in a representative operating environment, and prototype testing is an essential step in the transformation of research into industrial application.

#### *8.2.1.4 Additional Functionality*

Except for fault causes, other information is also useful to increase the levels of situational awareness, such as incipient fault locations and predicted failures. Therefore, some additional functions based on the proposed models can be investigated to further improve situational awareness in future. For example, the detected abnormal waveform can be further analyzed by an appropriate model to identify where these abnormal signals come from. And the waveform similarity based on identified recurrent faults can be further used to predict failure occurrences – if the sequence can be learnt from examples, then it can used as a predictive model.

## References


[1] J. A. Wischkaemper, C. L. Benner, B. D. Russell, and K. Manivannan, "Application of Waveform Analytics for Improved Situational Awareness of Electric Distribution Feeders," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2041–2049, 2015.

[2] K. M. Manivannan, "Power Distribution System Event Classification Using Fuzzy Logic," University of Texas A&M, 2012.

[3] W. Zhang, X. Xiao, K. Zhou, W. Xu, and Y. Jing, "Multicycle Incipient Fault Detection and Location for Medium Voltage Underground Cable," *IEEE Trans. Power Deliv.*, vol. 32, no. 3, pp. 1450–1459, 2017.

[4] T. S. Sidhu and Z. Xu, "Detection of incipient faults in distribution underground

cables," *IEEE Trans. Power Deliv.*, vol. 25, no. 3, pp. 1363–1371, 2010.

[5] P. A. Venikar, M. S. Ballal, B. S. Umre, and H. M. Suryawanshi, "Sensitive incipient inter-turn fault detection algorithm for power transformers," *IET Electr. Power Appl.*, vol. 10, no. 9, pp. 858–868, 2016.

[6] B. Li, Y. Jing, and W. Xu, "A Generic Waveform Abnormality Detection Method for Utility Equipment Condition Monitoring," *IEEE Trans. Power Deliv.*, vol. 32, no. 1, pp. 162–171, 2017.

[7] B. D. Russell and C. L. Benner, "Intelligent Systems for Improved Reliability and Failure Diagnosis in Distribution Systems," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 48–56, 2010.

[8] IEEE Working Group on Power Quality Data Analytics, "Electrical Signatures of Power Equipment Failures Electric Signature Analytics for Equipment Condition Monitoring," 2016. [Online]. Available: http://grouper.ieee.org/groups/td/pq/data/downloads/Signatures_Equipment_Failures_V2018Dec.pdf.

[9] Roy Hoffman, "Automated Distribution Feeder Fault Management: Possibilities and Challenges," *Electric Energy T&D Magazine*, Quebec, Canada, p. 44, 2009.

[10] the U.S. Department of Energy, "The Smart Grid: An Introduction."

[11] D. Xenias, C. J. Axon, L. Whitmarsh, P. M. Connor, N. Balta-ozkan, and A. Spence, "UK smart grid development : An expert assessment of the benefits, pitfalls and functions," *Renew. Energy*, vol. 81, pp. 89–102, 2015.

[12] K. Muthu-Manivannan, C. L. Benner, P. Xu., and B. D. Russell, "Electrical Power System Event Detection And Anticipation," US 8.457,910 B2, 2013.

[13] K. Muthu-Manivannan, C. L. Benner, P. Xu, and B. D. Russell, "Identification Of

Power System Events Using Fuzzy Logic," US 8,370.285 B2, 2013.

[14] UKPN and National Grid, "Transmission and Distribution Interface 2.0 (TDI)," 2017.

[15] J. Wu, Y. He, and N. Jenkins, "A robust state estimator for medium voltage distribution networks," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1008–1016, 2013.

[16] H. Liang, Y. Liu, G. Sheng, and X. Jiang, "Fault-cause identification method based on adaptive deep belief network and time – frequency characteristics of travelling wave," *IET Gener. Transm. Distrib.*, vol. 13, no. 5, pp. 724–732, 2019.

[17] X. Wang, "A Data Anlaytic Approach to Automatic Fault Diagnosis and Prognosis for Distribtion Automation," Univeristy of Strathclyde, 2017.

[18] Ofgem, "quality of service guaranteed standards." [Online]. Available: https://www.ofgem.gov.uk/licences-industry-codes-and-standards/standards/quality-service-guaranteed-standards. [Accessed: 02-Apr-2019].

[19] I. Hernando-Gil, I.-S. Ilie, and S. Z. Djokic, "Reliability Planning of Active Distribution Systems Incorporating Regulator Requirementfile:///D:/jojo/PhD/reading staff/Thesis/DNO annual report for award and penalty.pdfs and Network-Reliability Equivalents," *IET Gener. Transm. Distrib.*, vol. 10, no. 1, pp. 93–106, 2019.

[20] Ofgem, "RIIO-ED1 Annual Report 2015-16," 2016.

[21] J. Shin and H. Jun, "On condition based maintenance policy," *J. Comput. Des. Eng.*, vol. 2, pp. 119–127, 2015.

[22] Y. Ramesh, K. Santosh, V. A, L. A, and W. A, "Condition based maintenance of power transformer: A case study," in *2008 International Conference on Condition*

*Monitoring and Diagnosis*, 2008, pp. 502–504.

[23] M. Culhane, "Condition-based monitoring of oil- insulated , vacuum interrupting single-phase reclosers," 2017.

[24] X. Wang, S. D. J. Mcarthur, S. M. Strachan, J. D. Kirkwood, and B. Paisley, "A Data Analytic Approach to Automatic Fault Diagnosis and Prognosis for Distribution Automation," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6265–6273, 2018.

[25] X. Wang, "A Data Analytic Approach to Automatic Fault Diagnosis and Prognosis for Distribution Automation," University of Strathclyde, 2017.

[26] G. Burt, J. McDonald, A.G.King, J.Spiller, D.Brooke, and R.Samwell, "Intelligent online decision support for distribution system control and operation," *IEEE Trans. Power Syst.*, vol. 10, no. 4, pp. 1820–1827, 1995.

[27] Z. A. Vale, M. F. Fernandes, C. Rosado, C. Ramos, and L. Faria, "Better KBS for real-time applications in power system control centers : the experience of SPARSE project," *Comput. Ind.*, vol. 37, no. 2, pp. 97–111, 1998.

[28] P. Burrell and D. Inman, "An expert system for the analysis of faults in an electricity supply network: problems and achievements," in *Computers in Industry*, 1998, pp. 113–123.

[29] D. R. Sevcik, R. B. Lunsford, M. Kezunovic, Z. Galijasevic, S. Banu, and T. Popovic, "automated analysis of fault records and dissemination of event reports," in *Fault and Disurbance Analysis Conference*, 2000.

[30] M. Kezunovic and I. Rikalo, "Automating the analysis of faults and power quality," *EEE Comput. Appl. Power*, vol. 12, no. 1, pp. 46–50, 1999.

[31] S. D. J. Mcarthur and E. M. Davidson, "Automated Post-fault Diagnosis of Power System Disturbances," in *IEEE PES PSACE Panel Paper- General Meeting*, 2006.

[32] J. A. Hossack, S. Member, J. Menal, S. D. J. Mcarthur, and J. R. Mcdonald, "A Multiagent Architecture for Protection Engineering Diagnostic Assistance," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 639–647, 2003.

[33] G. Napier, E. M. Davidson, S. D. J. Mcarthur, S. Member, and J. R. Mcdonald, "An Automated Fault Analysis System for SP Energy Networks : Requirements , Design and Implementation," in *2009 IEEE Power & Energy Society General Meeting*, 2009, pp. 1–7.

[34] E. M. Davidson, S. D. J. Mcarthur, J. R. Mcdonald, S. Member, T. Cumming, and I. Watt, "Applying Multi-Agent System Technology in Practice : Automated Management and Analysis of SCADA and Digital Fault Recorder Data," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 559–567, 2006.

[35] B. M. Kezunovic, "Translational Knowledge : From Collecting Data to Making Decisions in a Smart Grid," *Proceeding IEEE*, vol. 99, no. 6, 2011.

[36] M. Moreto and J. G. Rolim, "Using disturbance records to automate the diagnosis of faults and operational procedures in power generators," *IET Gener. Transm. Distrib.*, vol. 9, pp. 2389–2397, 2015.

[37] X. Luo, S. Member, and M. Kezunovic, "Fault Analysis Based on Integration of Digital Relay and DFR Data," in *IEEE Power Engineering Society General Meeting, 2005*, 2005, p. 746–751 Vol. 1.

[38] J. Wright and G. Devine, "An innovative solution to monitoring 'Medium Voltage' Joints." [Online]. Available: https://www.eatechnology.com/wp-content/uploads/2017/08/Cable-Canary-for-CABLEx-2017.pdf. [Accessed: 26-Mar-2019].

[39] EA Technology, "UltraTEV Monitor." [Online]. Available:

https://www.eatechnology.com/americas/wp-content/uploads/sites/5/2017/03/UltraTEV-Monitor-Technical-Info-Brochure-US.pdf [Accessed: 15-June-2020]

[40] J. A. Wischkaemper and J. S. Bowers, "Distribution Fault Anticipation Improving Reliability and Operations by Knowing What Is Happening on Your Feeders," 2014. [Online]. Available: https://www.techadvantage.org/wp-content/uploads/2014/03/1B_Bowers-and-Wischkaemper.pdf. [Accessed: 27-Mar-2019].

[41] United Energy, "Distribution Fault Anticipation Data Collection and Analytics (DFADCAA)."[Online]. Available: https://www.aer.gov.au/system/files/United Energy - RRP 5-23 - DFADCAA PJ1599 - Jan 2016.pdf [Accessed: 15-June-2020]

[42] C. Manyame, R. E. Taylor, C. L. Benner, and D. Russell, "Mid-South Synergy Uses DFA Technology to Diagnose Fault-Induced Conductor Slap," 2017. [Online]. Available: https://www.lordconsulting.com/images/stories/pdf/20170611.MidSouth.ComplexFaultWithFics.pdf. [Accessed: 27-Mar-2019].

[43] EATON, "Blackout Tracker United Kingdom Annual Report 2015," 2015.

[44] K. L. Butler, B. D. Russell, C. Benner, and K. Andoh, "Characterization of Electrical Incipient Fault Signature Resulting from Tree Contact with Electric Distribution Feeders," in *IEEE Power Engineering Society Summer Meeting. Conference Proceedings*, 1999, pp. 408–413.

[45] M. Izadi *et al.*, "The influence of lightning induced voltage on the distribution power line polymer insulators," *PLoS One*, vol. 12, pp. 1–22, 2017.

[46] EPRI, "DOE/EPRI National Database Repository of Power System Events."

[Online]. Available: http://pqmon.epri.com/disturbance_library/.[Accessed: 23-Jun-2017].

[47] X. Qin, P. Wang, Y. Liu, L. Guo, and G. Sheng, "Research on Distribution Network Fault Recognition Method Based on Time-Frequency Characteristics of Fault Waveforms," *IEEE Access*, vol. 6, pp. 7291–7300, 2018.

[48] Transmission & Distribution Committee - IEEE Power & Energy Society, "Electric Signatures of Power Equipment Failures," no. May, 2015.

[49] L. A. Irwin, "Real Experience Using Power Quality Data to Improve Power Distribution Reliability," in *Proceedings of 14th International Conference on Harmonics and Quality of Power - ICHQP 2010*, 2010, pp. 1–4.

[50] B. Stephen, X. Jiang, and S. D. J. Mcarthur, "Extracting Distribution Network Fault Semantic Labels from Free Text Incident Tickets," *IEEE Trans. Power Deliv.*, 2019.

[51] C. Rudin *et al.*, "Machine Learning for the New York City Power Grid," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 328–345, 2012.

[52] J. A. Wischkaemper, "Electrical Characterization of Arcing Fault Behaviors on 120/208V Secondary Networks," Texas A&M University, 2011.

[53] B. Mak, "Basics of Load Switches," 2018. [Online]. Available: https://www.ti.com/lit/an/slva652a/slva652a.pdf?ts=1593766882652&ref_url=https%253A%252F%252Fwww.google.com%252F [Accessed: 15-June-2020]

[54] W. Zhang, Y. Jing, and X. Xiao, "Model-Based General Arcing Fault Detection in Medium-voltage Distribution Lines," *IEEE Trans. Power Deliv.*, vol. 31, no. 5, pp. 2231–2241, 2016.

[55] M. Kizilcay and P. La Seta, "Digital Simulation Of Fault Arcs In Medium-Voltage Distribution Network," in *15th Power Systems Computation Conference (PSCC'05)*, 2005,

no. August, pp. 22–26.

[56] Y. Tian and W. Xu, "A Method to Calculate Capacitor Switching Transients Using Short-Circuit Programs," in *16th International Conference on Harmonics and Quality of Power (ICHQP)*, 2014, no. 1, pp. 185–189.

[57] T. Ghanbari, E. Farjah, and F. Naseri, "Three-phase resistive capacitor switching transient limiter for mitigating power capacitor switching transients," *IET Gener. Transm. Distrib.*, vol. 10, pp. 142–153, 2016.

[58] C. Muscas, "Power Quality Monitoring in Modern Electric Distribution Systems," *IEEE Instrum. Meas. Mag.*, no. October, 2010.

[59] S. P. Verma and P. Kumar, "Smart Grid , Its Power Quality and Electromagnetic Compatibility," *MIT Int. J. Electr. Instrum. Eng.*, vol. 2, no. 1, pp. 55–64, 2012.

[60] K. D. Mcbee, M. G. Simões, and S. Member, "Evaluating the Long-Term Impact of a Continuously Increasing Harmonic Demand on Feeder-Level Voltage Distortion," *IEEE Trans. Ind. Appl.*, vol. 50, no. 3, pp. 2142–2149, 2014.

[61] C. Venkatesh, D. S. Kumar, D. V. S. S. S. Sarma, and M. Sydulu, "Modelling of Nonlinear Loads and Estimation of Harmonics in Industrial Distribution System," in *Fifteenth National Power Systems Conference (NPSC)*, 2008, vol. 2, no. December, pp. 592–597.

[62] S. M. Adnaan and M. Tech, "Design and Simulation of a nonlinear Load Model Used to Simulate Voltage Notches and Harmonics Caused by a 6-Pulse Three-Phase Rectifier," *Int. J. Sci. Eng. Technol. Res.*, vol. 6, no. 6, pp. 984–995, 2017.

[63] K. P. Schneider *et al.*, "Analytic Considerations and Design Basis for the IEEE Distribution Test Feeders Report Prepared by the Test Feeder Working Group of the Distribution System Analysis Subcommittee," *IEEE Trans. Power Syst.*, vol. 33,

no. 3, pp. 3181–3188, 2018.

[64] IEEE PSRC Working Group D15, "High Impedance Fault Detection Technology," 1996.[Online]. Available: https://www.pes-psrc.org/kb/published/reports/High_Impedance_Fault_Detection_Technology.pdf [Accessed: 15-June-2020]

[65] Russell Bryans, "Calculation of system fault levels," 2017. [Online]. Available: https://www.spenergynetworks.co.uk/userfiles/file/ESDD-02-006.pdf.

[66] IEEE Working Group on Power Quality Data Analytics, "2 day gap-less waveform data." [Online]. Available: http://grouper.ieee.org/groups/td/pq/data/. [Accessed: 21-Feb-2019].

[67] Transmission and Distribution Committee, "IEEE Std 1159™-2009, IEEE Recommended Practice for Monitoring Electric Power Quality," p. 53, 2009.

[68] W. D. Penny and S. J. Roberts, "Dynamic Models for Nonstationary Signal Segmentation," *Comput. Biomed. Res.*, vol. 502, pp. 483–502, 1999.

[69] G. D. Antona, C. Muscas, and S. Sulis, "Localization of Nonlinear Loads in Electric Systems Through Harmonic Source Estimation," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 10, pp. 3423–3430, 2011.

[70] Transmission & Distribution Committee, Power Quality Subcommittee, and IEEE Working Group on Power Quality Data Analytics, "Electric Signatures of Power Equipment Failures," 2018. [Online]. Available: http://grouper.ieee.org/groups/td/pq/data/downloads/Signatures_Equipment_Failures_V2018Dec.pdf.

[71] R. D. Telford, S. Galloway, B. Stephen, and I. Elders, "Diagnosis of Series DC Arc Faults—A Machine Learning Approach," *IEEE Trans. Ind. Informatics*, vol. 13, no.

4, pp. 1598–1609, 2017.

[72] E. Gulski, "Digital Analysis of Partial Discharges," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 2, no. 5, 1995.

[73] R. P. Adams and D. J. C. MacKay, "Bayesian Online Changepoint Detection," 2007.

[74] P. Fearnhead, "Exact and efficient Bayesian inference for multiple changepoint problems," *Stat. Comput.*, vol. 16, no. 2, pp. 203–213, 2006.

[75] S. Maleki, C. Bingham, and Y. Zhang, "Development and Realization of Changepoint Analysis for the Detection of Emerging Faults on Industrial Systems," *IEEE Trans. Ind. Informatics*, vol. 12, no. 3, pp. 1180–1187, 2016.

[76] P. Fearnhead, "On-line inference for hidden Markov models via particle filters," *J. R. Stat. Soc.*, pp. 887–899, 2003.

[77] U. Dwivedi and S. Singh, "Denoising Techniques With Change-Point Approach for Wavelet-Based Power-Quality Monitoring," *IEEE Trans. Power Deliv.*, vol. 24, no. 3, pp. 1719–1727, 2009.

[78] B. W. Yap and C. H. Sim, "Comparisons of various types of normality tests," *J. Stat. Comput. Simul.*, vol. 9655, no. May, pp. 2141–2155, 2011.

[79] A. Loy, L. Follett, H. Hofmann, and M. E. Mar, "Variations of Q-Q Plots – The Power of our Eyes !," *Am. Stat.*, vol. 54911, pp. 1–21, 2014.

[80] D. Fink, "A Compendium of Conjugate Priors," 1997. [Online]. Available: https://www.johndcook.com/CompendiumOfConjugatePriors.pdf [Accessed: 21-June-2020]

[81] K. P. Murphy, "Conjugate Bayesian Analysis of the Gaussian Distribution," *Def*, vol. 1, no. 7, pp. 1–29, 2007.

[82] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. 2012.

[83] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.

[84] L. Xu and M.-Y. Chow, "A Classification Approach for Power Distribution Systems Fault Cause Identification," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 53–60, 2006.

[85] L. Xu, M. Chow, and L. S. Taylor, "Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E - Algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 164–171, 2007.

[86] U. J. Minnaar, F. Nicolls, and C. T. Gaunt, "Automating Transmission-Line Fault Root Cause Analysis," *IEEE Trans. Power Deliv.*, vol. 31, no. 4, pp. 1692–1700, 2016.

[87] C. L. Benner and B. D. Russell, "Feeder Interruptions Caused by Recurring Faults on Distribution Feeders : Faults You Don't Know About," in *2008 61st Annual Conference for Protective Relay Engineers*, 2008, pp. 584–590.

[88] K. Manivinnan, C. L. Benner, B. D. Russell, and J. A. Wischkaemper, "Automatic Identification , Clustering and Reporting of Recurrent Faults in Electric Distribution Feeders," in *Intelligent System Application to Power Systems (ISAP)*, 2017.

[89] U. J. Minnaar, C. T. Gaunt, and F. Nicolls, "Characterisation of power system events on South African transmission power lines," *Electr. Power Syst. Res.*, vol. 88, pp. 25–32, 2012.

[90] S. Hiroaki and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoust.*, vol. 26, no. 1, 1978.

[91] M. Chow and L. S. Taylor, "Analysis And Prevention Of Animal-Caused Faults In Power Distribution Systems," *IEEE Trans. Power Deliv.*, vol. 10, no. 2, pp. 995–1001, 1995.

[92] R. A. Rossi, N. K. Ahmed, H. Eldardiry, and R. Zhou, "Similarity-based Multi-label Learning," in *arxiv*, 2017.

[93] M. Chow, S. Yee, and L. S. Taylor, "Recognizing Animal-Caused Faults in Power Distribution Systems Using Artificial Neural Networks," *IEEE Trans. Power Deliv.*, vol. 8, no. 3, pp. 1268–1274, 1993.

[94] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 1998.

[95] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning-data mining,inference, and prediction*. Springer Series in Statistics, 2017.

[96] R. J. Passonneau, C. Rudin, A. Radeva, and Z. A. Liu, "Reducing Noise in Labels and Features for a Real World Dataset: Application of NLP Corpus Annotation Methods," in *Computational Linguistics and Intelligent Text Processing*, 2009, pp. 86–97.

[97] C. Xie, G. Zou, H. Wang, and Y. Jin, "A new condition assessment method for distribution transformers based on operation data and record text mining technique," in *2016 China International Conference on Electricity Distribution*, 2016, no. Ciced, pp. 10–13.

[98] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[99] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[100] L. Yao, D. Mimno, and A. Mccallum, "Efficient Methods for Topic Model Inference on Streaming Document Collections," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[101] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," vol. 18, no. 11, 1975.

[102] X. Jiang, B. Stephen, and S. D. J. . McArthur, "Automated Fault Analysis and Diagnosis using High-Frequency and Maintenance Data from Distribution Networks," in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, 2019, pp. 1–5.

[103] X. Jiang, B. Stephen, and S. D. J. Mcarthur, "A Sequential Bayesian Approach to Online Power Quality Anomaly Segmentation," *IEEE Trans. Ind. Informatics*, pp. 1–10, 2020.

[104] X. Jiang, B. Stephen, and S. D. J. Mcarthur, "Automated Distribution Network Fault Cause Identification with Advanced Similarity Metrics," *IEEE Trans. Power Deliv.*, pp. 1–8, 2020.