Automated Distribution Network Fault Cause Identification with Advanced Similarity Metrics

Xu Jiang, *Student Member IEEE,* Bruce Stephen, *Senior Member IEEE,* Stephen McArthur, *Fellow IEEE*

Abstract— **Distribution network monitoring has the potential to improve service levels by reporting the origin of fault events and informing the nature of remedial action. To achieve this practically, intelligent systems to automatically recognize the cause of network faults could provide a data driven solution, however, these usually require a large amount of examples to learn from, making their implementation burdensome. Furthermore, the choice of input to such a system in order to make accurate classifications is not always clear. In response to this challenge, this paper contributes a means of using minimal amounts of historical fault data to infer fault cause from substation current data through a novel structural similarity metric applied to the associated power quality waveform. This approach is demonstrated along with disturbance context similarity assessment on an industrially relevant benchmark data set where it is shown to provide an improvement in classification accuracy over comparable techniques.**

Index Terms—**Fault Cause Diagnostic, Waveform Similarity, Context Similarity, Distribution Networks**

I. Introduction

The increasing complexity of distribution networks coupled with their limited observability can prolong unplanned outages from faults. This is highly undesirable from a customer perspective, and in addition the network operator will receive regulatory penalties based on the number of customer minutes being off supply. The situation is further complicated through the integration of low carbon technologies with legacy plant. This results in new and previously unconsidered faults, which may have unfamiliar characteristics when observed operationally.

Traditionally, fault causes were identified through manual analysis of weather and fault behavior [1]. The expert knowledge that defines this is difficult to standardize across cases and, as a result, fault cause identification is time-consuming and therefore expensive to undertake. Additionally, the complex form faults can now take makes this endeavor more challenging, as the existing knowledge does not extend to the new fault types. High-resolution fault and disturbance recording equipment is increasingly being implemented to support fault analysis, but it compounds the problem further, in that the waveform level representations they capture are too voluminous to interpret manually.

In response to this, recent research has considered using automatic classifiers: [2] 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. Other research proposed Artificial Neural Network (ANN) [3] and Fuzzy Classification [4] using field context data to identify outage causes via training with a large amount of examples. [5] utilized One Nearest Neighbor (1-NN) to rank and validate the relevant contextual and waveform features for transmission level fault identification, while [1] constructed a Deep Belief Network (DBN) to identify fault cause. Many state of the art classifiers [1] [3] [4] [5] require thousands of examples to learn from which is time consuming and impractical. Despite the potential operational benefit, most utility companies would consider archiving curated fault data marked up with diagnostic labels to be beyond their usual remit. However, previous research [6][7][8] identified that many faults and failures can exhibit similar characteristics. This would be classed as “event similarity”.

Event similarity could be used to automatically identify a recurring fault situation via patterns learned from this historical data [9], which can in turn be used for diagnosis and prognosis of recurrent incipient faults observed operationally [10][11]. Operational noise and variability make matching up events with historical equivalents difficult, necessitating means of similarity to be developed specifically for waveforms.

To support the application of a fault cause identifier for practical use on distribution networks, the following problems need to be addressed: extensive labeled fault examples are not always available for training classifiers, therefore, this paper proposes a means of inferring fault cause from operational data through analyzing the most similar Power Quality (PQ) events on a distribution network; fault signatures can vary in duration and magnitude even when they result from the same cause [9][12] - 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. Operationally, the resulting classifier can simply be embedded into an existing control center fault reporting process and propagate the predicted fault context to maintenance crews who in turn can approach root cause investigations with higher situational awareness. Synthesized fault data from physics based simulations may lack the realistic variability that operational data will exhibit, so

X. Jiang, B.Stephen, and S. D. J. McArthur are with the Institute for Energy and Environment, Department of Electronic and Electrical Engineering, University of Strathclyde, UK. (e-mail: x.jiang@strath.ac.uk, bruce.stephen@strath.ac.uk, s.mcarthur@strath.ac.uk).

testing must be undertaken on an operational data set to demonstrate effectiveness. The model capability and performance here are demonstrated on the US Department of Energy (DoE) Power Quality data set [13] which contains 334 three-phase voltage and current signals of PQ disturbances collected from an operational substation by the Electric Power Research Institute (EPRI).

The main contributions from this research are: 1) a new similarity metric which can identify recurring faults; 2) a similarity-based means of learning and automatically identifying fault causes which do not require waveform characteristic extraction; 3) a resulting classifier which does not require a large number of exemplars to learn from, which readily permits practical implementation; 4) a demonstration of successful fault classification with the proposed method. This performance is compared with conventional classifiers drawn from recent literature. Greater classification accuracy is demonstrated through combining waveform characteristics with fault context. 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 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 cause whereas the bird strike is transient in nature and therefore pointless to look for – hence restoration of power can be immediate. In practice, this would allow the circumstances in which faults occurred to be automatically diagnosed without domain expert intervention, leading to shorter investigation periods, pre-emption of faults at the incipient stage and, overall, shorter unplanned outages.

II. PQ Disturbance Data

PQ disturbance causes are multifactorial which presents difficulties in identifying features that represent particular faults [14][5]. The DoE PQ data set provides 166 expert labeled three-phase AC voltage and current signals sampled at 0.96 and 3.84 kHz [13]; one such event is shown in Fig. 1, which is a short-term single-phase to ground fault on both the voltage and current signals attributed to an overhead arrester failure. The amplitude shift starts at approximately 0.002s and ends at 0.044s with a re-closer operation, suggesting that the fault is probably not eliminated entirely. Fig. 1 highlights that the fault changes the waveform shape of more than just one phase and not just in terms of its magnitude or relation to other phase waveforms. Additionally, it also provides the fault waveform start time and end time down to the millisecond level, associated weather, fault cause and associated isolation equipment.

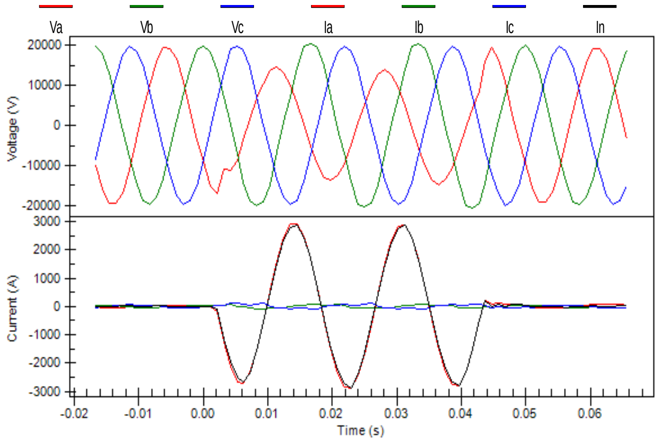

Fig. 1. Power Quality Waveforms for short term phase-earth overcurrent. The fault clears in 0.042 sec; overhead arrester failure; isolated by recloser; clear weather; happened at 5/19/2005 04:40:26.1990, Phase A

A. Discriminatory Features

In prior research, statistical and signal features have been used to distinguish the cause of PQ disturbances [2][13]. Despite this, it can still be unclear which features are appropriate to assess fault cause especially when minimal exemplars are available. To illustrate the potential inseparability of the relative phase faults, a visualization of the phase distribution of the DoE PQ library signals is shown in Fig. 2. Absolute current magnitude is insufficient to describe phase faults, because the same magnitude under different voltage levels provides different waveform attributes. An appropriate and intuitive visualization for the relative values between phases are compositional techniques[14][15], which can visualize the proportion of current and symmetrical components taken on different phases as a 3-Simplex. The symmetrical components of current signals have been used to classify fault types previously[16]. The symmetrical components of current signals can be expressed as:

$$I_p = \frac{1}{3}(I_A + aI_B + a^2 I_C) \tag{1}$$

$$I_n = \frac{1}{3}(I_A + a^2 I_B + aI_C) \tag{2}$$

$$I_z = \frac{1}{3}(I_A + I_B + I_C) \tag{3}$$

where I_p, I_n, I_z respectively represent positive, negative and zero sequence current, I_A, I_B, I_C represent three phase current and a is defined as a phase rotation, which rotates a phasor vector courter clockwise by 120 degrees. The legend in Fig. 2 separates faults according to their causes (tree, equipment, vehicle, animal contact, lightning strike) which are included in the EPRI DoE dataset. As Fig. 2 illustrates, different faults have a more identifiable, but still unclear, boundary that is dependent on their cause. The positive component is dominant, forcing all points into a corner of the simplex; the

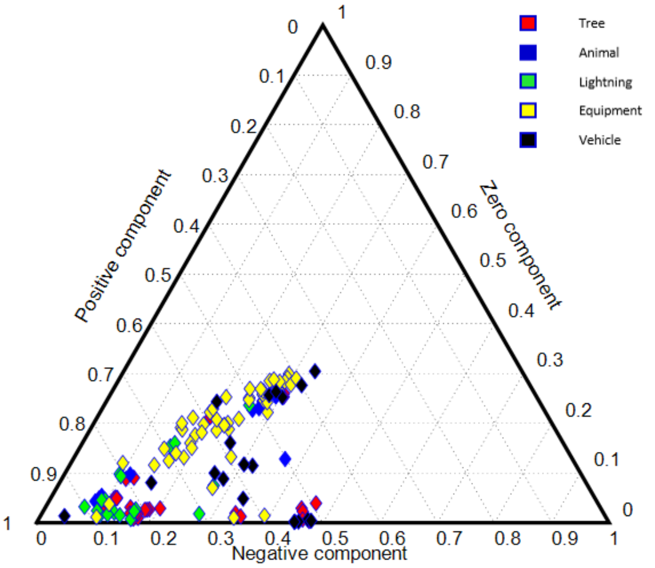

Fig. 2. Phase symmetrical components representation of PQ faults annotated with causes. This representation, although popular, offers little to distinguish fault by type.

contributions of the negative and zero component, while reflective of waveform asymmetry and earthing faults, are still insufficiently discriminative to distinguish these from other faults. Therefore, it is worthwhile investigating methods to enhance discriminatory power by using other features. Calculating similarity of events can meet both requirements.

III. PQ Disturbance Similarity

The Nearest Neighbor (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 [5]. 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. 3 illustrates the processing stages of the proposed similarity-based classification model. From Fig. 3, the waveform processing output stages are associated with context obtained from fault recorders, such as isolation equipment operated, and weather predicates such as localized environmental conditions. Since faults manifest as abrupt noise signals rather than changes in periodicity, the noise 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 closet historical event and infer the associated fault cause for reporting. The detailed function of these processing stages will now be described.

A. *Waveform Pre-processing*

To mitigate the influence of the sinusoidal waveform on the similarity metric, the fault components can be extracted by removing the sinusoidal component. The conventional approach to decoupling the sinusoidal components from the abnormal components of the signal would be to superimpose faults onto the last normal cycle waveform [15], which can be simply expressed as:

$$f(t) = i(t) - i(t - qN_0\Delta t) \qquad (1)$$

where $f(t)$ represents the fault component at time t , q denotes the number of gap cycles between the last healthy cycle and the measured cycle. N_0 denotes the number of samples in one cycle of current; Δt is the time gap between two consecutive samples. This method utilizes the present measure superimposed over the last healthy cycle, which allows the shape of the residual fault components to be used to evaluate the waveform similarity. An example of the residual fault component is given in Fig. 4. As Fig. 4 shows, some faults, such as arcing, are usually triggered at the peaks [16]. When they initiate at peak or valley positions it affects the sign of the residual. To solve this, the absolute value of the fault components is used to evaluate the similarity between pairs.

B. *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 [17], 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:

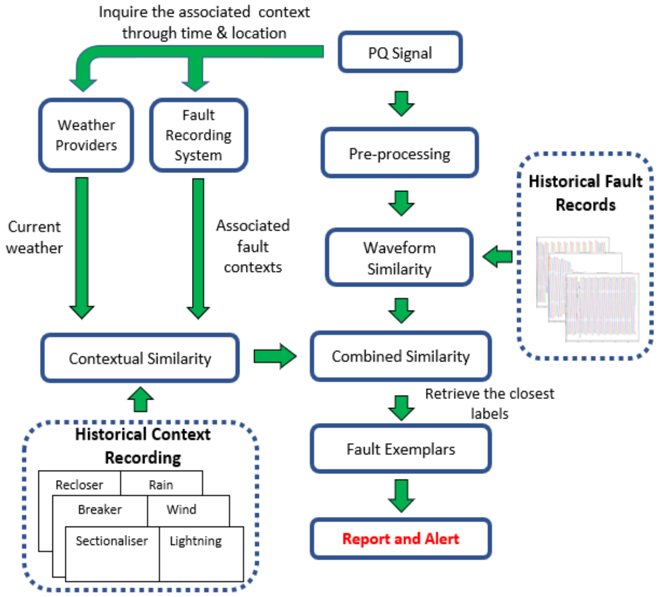

Fig. 3. Processing stages for both the training and testing phase of the proposed automated PQ disturbance classifier.

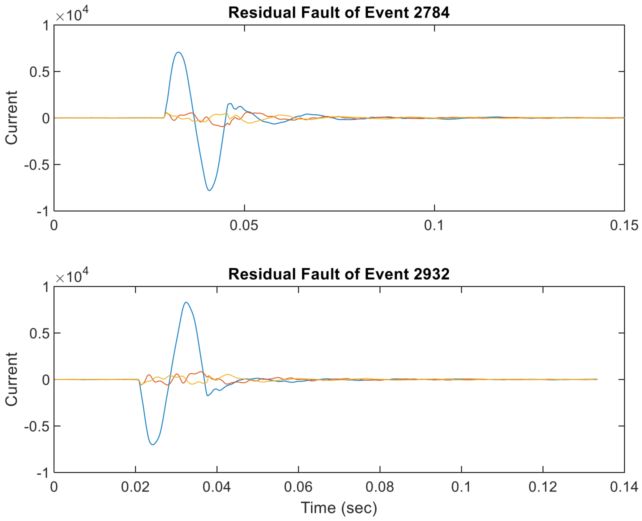

Fig 4 Residual fault components of event 2784 and event 2932

$$U = u_1, u_2 \dots u_n \dots u_N \qquad (2)$$

$$V = v_1, v_2 \dots v_m \dots v_M \qquad (3)$$

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 and provides an optimum path from the bottom left to the top right which is called the warping path, $WP(k)$, traverses as:

$$\begin{gathered} WP(k) = w(1), w(2) \dots w(k) \dots w(K), \\ \max(N, M) \leq K < N + M \end{gathered} \qquad (4)$$

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

$$C_k = \sum_{j=1}^{k} Dist(w(j)) \qquad (5)$$

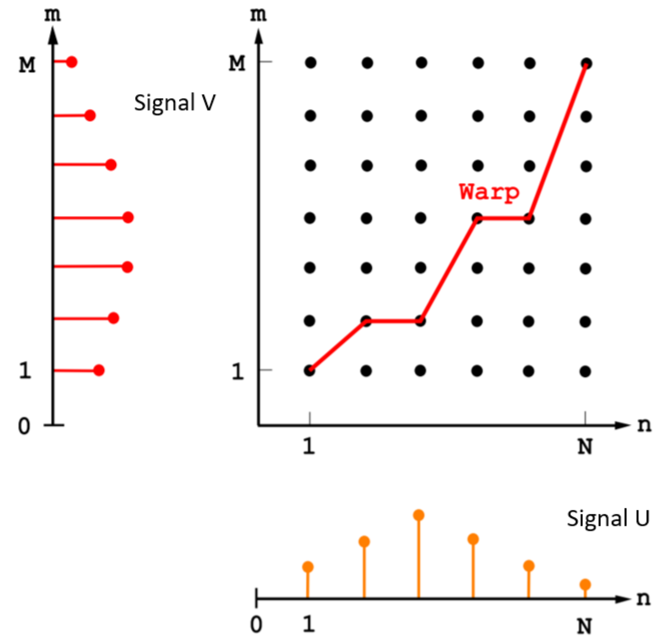

Fig. 5. DTW cost matrix formation for two signals Y and X of duration M and N; the warping path is defined as the lowest cost route from cell 1,1 to N, M.

$$D(WP(k)) = Min(C_k) \tag{6}$$

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

As Fig. 5 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:

$$\begin{aligned} D\big(WP(k+1)\big) = D\big(WP(k)\big) \\ + \min\,(Dist(i,j+1), Dist(i \\ +1, j+1), Dist(i+1,j)) \end{aligned} \tag{7}$$

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.

C. Contextual Similarity Measurement

With the context of the new fault extracted, contextual similarity can be evaluated. As Fig. 3 shows, the context can be extracted through time and location. This paper utilizes the same context data as in [3][4][20], 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 [18] :

TABLE I

CONTEXTUAL FEATURES USED FOR FAULT CAUSE IDENTIFICATION [3] [4] [20]

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

$$D_{H(U,V)} = \frac{1}{N_c}\sum_{i=1}^{N_c} |Y_{i(U)} - Y_{i(V)}| \tag{8}$$

where $Y_{i(U)}$, $Y_{i(V)}$ are the categories that represent the context of signal U and V respectively. N_c is the number of contextual features. The output of the Hamming distance is a discrete value. Additionally, timestamp is a continuous value which can be discretized into daytime and season labels using predefined ranges, which are shown in Table I.

D. Combined Similarity

Faults can manifest through their waveforms 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 that waveform similarity and contextual similarity are combined as follows:

$$Comb_{(U,V)} = \frac{D_{H(U,V)}}{\max\,(D_H{}')} \cdot \frac{D(WP(K))_{(U,V)}}{\max\,(D\big(WP(K)\big)')} \tag{9}$$

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

IV. 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 [13]. 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 experiments are presented to highlight the practical effectiveness of the metric: the first experiment validates that the proposed waveform similarity measurement can identify shape-based recurrent faults. The second experiment is to validate that the proposed contextual similarity can identify recurrent faults based on context. Both experiments use the same pair of events, shown in Fig. 6, for comparison purposes; the residual fault component of these was given in Fig. 4. The time interval between these two events in Fig. 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

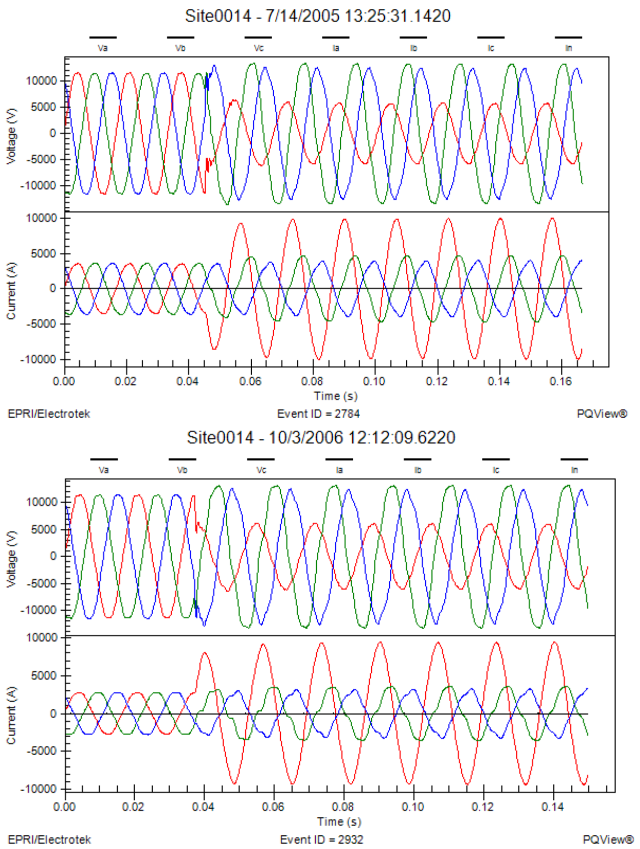

Fig. 6. two PQ disturbance events with a similarity approaching the maximum value. Although the cause is the same in both cases, a pairwise comparison would have overlooked this due to differences in the duration and cycle position of fault initiation.

TABLE II
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	

domain knowledge from facilitating fault analysis and therefore resolution. However, these two faults came from the same substation (Site 14 in the DoE set), and resulted from a terminator failure. The fault was interrupted by a circuit breaker. Consequently, the PQ waveform of these two events is very similar although from Fig. 6 it can be seen that event 2932 begins half a cycle earlier than event 2784. It is quite common to see incipient faults, such as arc faults in underground cables [8], begin at different parts of the cycle, which would have rendered a pointwise similarity metric ineffective.

A. Shape-based Signal Similarity

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

B. Contextual Similarity

The fault records contain associated information which is detailed in Table II. 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 parts of North America may be the same in Fall, Spring or Winter so this could be uninformative. For this case, the proposed method has calculated the contextual similarity between the two events in Fig. 6 as 0.8.

V. AUTOMATED FAULT CAUSE IDENTIFICATION BENCHMARKS

Using supervised classifiers to automatically identify fault cause [19] still requires domain knowledge to select appropriate input features. Previous research used two broad categories of features to identify fault causes in distribution networks: Waveform-based features [2] and contextual features [3] [4] [20]. The waveform-based features arise from field experience, for example, animal contact is likely to only affect a single phase owing to the nature 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. The features that were chosen using domain knowledge then extracted from the DoE data paper are shown in Table III and have been previously discussed in [2]. Other prior work [3] [4] [20] has incorporated fault context, such as weather, season, faulted phase and time of day to identify the fault cause. Examples of this have been listed in Table I of Section IIIC. To demonstrate the performance benefits of the proposed similarity measure in a classifier, it will now be benchmarked on operational data against the existing

TABLE III
WAVEFORM CHARACTERISTICS USED FOR FAULT CAUSE IDENTIFICATION

Symbol	Equation	Description
R_1	$\dfrac{\max\,(I_{amax}, I_{bmax}, I_{cmax})}{\text{median}\,(I_{amax}, I_{bmax}, I_{cmax})}$	Ratio used in logical expression to infer the number of faulted phases
R_2	$\dfrac{\text{median}\,(I_{amax}, I_{bmax}, I_{cmax})}{\min\,(I_{amax}, I_{bmax}, I_{cmax})}$	Ratio used in logical expression to infer the number of faulted phases
I_f	$I_{pkmax} - I_{0pk}$	Fault current component - fault peak value minus normal operational peak value
n_f	$\sum_{j=1}^{n}(\frac{I_{pk(j)}}{I_{0pk}} > R_{th})$	Fault duration - cumulative cycles where ratio exceeds a threshold R_{th}
α_{ATT}	$\dfrac{I_{pkmin}}{I_{0pk}}$	Fault current attenuation - the ratio of minimum peak three-phase value to normal operational peak
E	Energy(wavedec(Vnorm))	Frequency domain energy percentage – wavelet transform to inform the energy contribution of frequency bands

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: the peak current of the first cycle
I_{pkmin}- minimum value of peak point of phase current

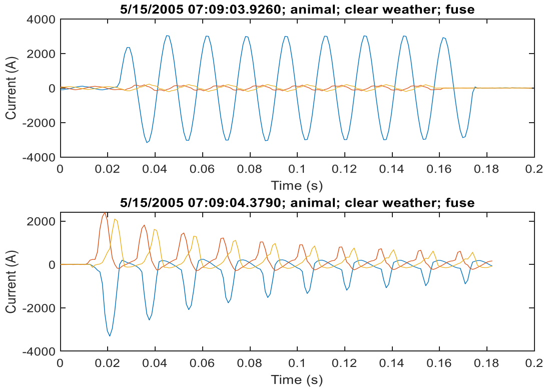

Fig. 7. Two consecutive animal fault episodes occurring less than a second apart; the waveform is dissimilar but context matches exactly

classifiers [1][3][5] with both sets of input features: waveform and context in order to show the benefit of using an advanced similarity measure over the conventional ones previously used.

VI. Automated Fault Cause Identification

The analysis in Section V has discussed the previous 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 training data and no previous work identifies an appropriate classifier that can work using minimal exemplars. To address this, this 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 equivalent models with waveform-based features, contextual features as well as a combination of both. No user tunable parameters are required. 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 examples 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 knowledge-based and statistical 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. The classifiers tested are ANN[2], DBN[1], Decision Tree, Discriminant, SVM, KNN[5] and Ensemble methods[21][22]. Furthermore, two common evaluation metrics, Overall Accuracy (ACC) and F-score are used to evaluate the performance [5]. 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.

A. Benchmark performance for feature based methodology

Past works [2][3] used waveform-based features and contextual features respectively to identify fault causes in distribution networks. Table IV shows the performance of different benchmark classifiers with both feature sets as well as the combination of the two. Regardless of the classifier chosen, the rank of the accuracy metrics show that the contextual features perform better than waveform-based features alone, but worse than the combination of the two. These will now be described.

1) Waveform Characteristics

Although 1-NN and Bagged Tree can identify fault cause to a reasonable level (> 60% in Table IV), 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 given in Fig. 7. The events in Fig. 7 recur consecutively within a short period and they were 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 the DoE data set all occur around April to August and frequently occur under fair weather, which means the contextual feature can be a more powerful predictor than waveform on these

TABLE IV
Comparison of Choice of Model And Feature set for Benchmark Fault Cause Classifier

Classifier	Feature Set	F-score					Overall Accuracy
		Tree	Equipment	Animal	Vehicle	Lightning	
ANN [3]	Waveform Features [2]	18.42%	32.76%	0%	8.33%	19.27	19.27%
Bagged Tree	Waveform Features [2]	67.47%	79.49%	40%	66.67%	52.94%	69.88%
1-NN [5]	Waveform Features [2]	67.42%	66.67%	20%	60.47%	48.48%	61.44%
ANN [3]	Contextual Features [3] [4] [20]	22.86%	28.8%	10.81%	22.64%	12.77%	22.22%
Bagged Tree	Contextual Features [3] [4] [20]	78.57%	78.67%	60%	**88.37%**	75.68%	78.9%
1-NN [5]	Contextual Features [3] [4] [20]	65.71%	82.67%	76.19%	86.38%	63.86%	76.6%
ANN [3]	Combined Features [3] [2]	22.78%	39.62%	21.26%	21.05%	14.04%	24.09%
DBN [1]	Combined Features [3] [2]	0%	62.24%	0%	0%	0%	43.43%
Bagged Tree	Combined Features [3] [2]	77.92%	82.72%	**73.68%**	85%	82.35%	81.32%
1-NN [5]	Combined Features [3] [2]	**80.95%**	**86.9%**	66.67%	79.07%	**83.33%**	**82.5%**

TABLE V
Comparison of Choice of Similarity Measure for Proposed 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%**

faults. Generally, the waveform characteristics alone can be difficult to use to generalize fault causes, especially for faults with significant variability.

2) Contextual Features

Compared to waveform-based classification, the use of contextual features alone can improve the overall accuracy by approximately 10% for both 1-NN (from 61.44% to 76.66%) and Bagged Tree (from 69.88% to 78.9%) as Table IV shows. Generally, contextual features are considered powerful predictors of fault causes [3], however, the accuracy for contextual features alone is not enough in practical implementation as Table IV shows.

3) Combined Features

Table IV shows that the existing classifiers with combined waveform derived inputs, augmented with features based on context, generally outperform the classifiers that that only used the individual feature sets; e.g. 1-NN, Bagged Tree and ANN achieve 82.5%, 81.32% and 24.09% which improve approximately 6%, 2% and 2% respectively. The 0% accuracy obtained by DBN shows it cannot work with a minimal number of exemplars. This shows that waveform and context together carries additional information to support accurate classification.

B. Performance comparison with similarity-based methodology

To investigate the predictive power of the proposed similarity measures against conventional waveform-based features and contextual features, three comparison experiments are carried out using the highly performing 1-NN as the classification model. Table V shows the performance for classifiers using the proposed similarity measures. An overall ACC (>60%) is obtained using these with a 1-NN classifier. Incorporating the proposed similarity measure results in improvement in all event type classifications over an equivalent classifier in Table IV that used using Euclidean distance on conventional waveform features. Every fault class attains an F-score of greater than 75% and lightning strike classification accuracies are improved by almost 11% to 94.12%. The overall best classification accuracy, 89.15%, comes from using the combined similarity measure, which improves accuracy by approximately 7% over conventional combined input features and without the need to pre-process any waveform statistical features. Generally, the advanced similarity based methodology can be advantageous regardless of waveform or contextual inputs used. Fault diagnosis using combined similarity still can achieve higher accuracy than using either waveform or context alone.

C. Consequence of Minimal Data Support on Classifiers Performance

The intended benefit of the proposed approach is that it will require the minimum number of examples to learn from. Instead of waiting to collect a large number of exemplars, utilities will inevitably prefer an intelligent classifier that works with minimal available data in order to gain value from monitoring as quickly as possible. As Table IV And Table V show, regardless of the features chosen, Neural Networks, including ANN and DBN, cannot classify fault cause with minimal support, with ANN only managing 24.09% overall accuracy and DBN achieving 43.43% overall accuracy, in both cases using the combined features. 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.

TABLE VI
COMPARISON OF WAVEFORM SAMPLING FREQUENCIES

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

D. Performance Impact of Sampling Frequency

In practice, different sampling frequencies have been used to record PQ events, including 960Hz and 3840Hz [13][23]. For conventional fault classification, the sampling frequency can affect the waveform feature extraction then further affect the classification accuracy. To demonstrate, the best performing 1-NN classifier from Table V were run again using input data with the aforementioned sampling frequencies in separate groups. Table VI shows the separation of these results: the events recorded at higher sampling frequency can achieve 94.44% overall classification accuracy – an almost 20% gain which justifies the higher resolution of the data. This performance benefit had been obscured in Table V by the aggregation of both data resolutions which had led to an accuracy of 89.15%.

E. Interpretation of Fault Classification Performance

The comparison of the classifiers and input feature sets in this section resulted in a range of classification accuracies. These can be attributed two factors: the ability of the model to form a sufficiently expressive decision boundary, and the ability of the inputs to provide discriminatory power between fault types. Erosion of this accuracy can be attributed to the heightened variability of particular fault cases. Consequently, waveform classification on its own is generally poor for animal related fault causes – it is difficult for a classifier to capture a general representation from the number of combinations of type of animal, circumstances and equipment affected. Vehicle impacts are similarly varied and lead to lower classification performance as a result: contact angle, size and speed of vehicle and span geometry will all contribute to how an overhead line pole impact manifests on a PQ waveform in terms of phase affected and event duration. In contrast, lightning related fault episodes are distinct, which is understandable given the unique high energy nature of the fault and the consistently short timescales over which strikes tend to occur.

VII. CONCLUSIONS

Power distribution networks are featuring greater levels of observability given the availability of lower cost sensor and monitoring systems. The value of these to enhancing power delivery service levels can only be realized if data can be interpreted in an automated and repeatable manner. This paper has contributed a new similarity measure to identify recurrent faults and built a classification methodology to automatically identify fault causes associated with Power Quality events from a minimal number of exemplars. The highest accuracy achieved with the contributed approach is 89.15%, using a combination of waveform and contextual similarity and retaining high accuracy even for low-prevalence events. As a utility maintaining distribution network assets, the ability to identify the causes of disturbances via Power Quality 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 at risk assets to be planned if the frequency of occurrence (and possibly the time of year) is known. Without a means of converting fault waveforms into meaningful representations, this actionable insight would not be available. Moving towards an operational solution for the works contributed here, supporting research is now required to understand the heterogeneity of fault waveforms invoked by network topology, materials and design of equipment used (e.g. insulation in switchgear) and influence of seasonal or diurnal effects. It is envisaged that in order for this to be realized, a hardware based implementation is required to form the basis of a pilot study based around small scale operational deployment on a network with well understood issues.

VIII. Reference

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Y. Song, W. Wang, Z. Zhang, H. Qi, and Y. Liu, “Multiple event analysis for large-scale power systems through cluster-based sparse coding,” *Trans. Power Syst.*, vol. 32, no. 6, pp. 301–306, 2017.

[13] EPRI, “DOE/EPRI National Database Repository of Power System Events.” [Online]. Available: http://pqmon.epri.com/disturbance_library/.[Accessed: 23-Jun-2017].

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] 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.

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

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

[23] Transmission and Distribution Committee, “IEEE Std 1159™-2009, IEEE Recommended Practice for Monitoring Electric Power Quality,” p. 35, 2009.