Automating Transmission Line Fault Root Cause Analysis

UJ Minnaar*, F Nicolls**, and CT Gaunt**, *Member, IEEE*

*Eskom Holdings, South Africa, ** Department of Electrical Engineering, University of Cape Town, South Africa*

***Abstract*— This research demonstrates that transmission line faults can be classified automatically according to their underlying cause, and lays a foundation for operational classification of transmission line faults in system control centres. The transmission line fault waveforms are characterised by instantaneous symmetrical component analysis to describe the transient and steady state fault conditions. Using a large fault record and waveform database, classification features based on the waveform and external environmental characteristics have been identified to develop single-nearest-neighbour classifiers that identify the underlying cause of transmission line faults, and good classification accuracy has been achieved.**

***Index Terms*— power system operations, transmission line faults, fault causes, nearest neighbor classifier, pattern recognition**

I. Introduction

MODERN society depends on electricity supplies that are reliable [1] and compatible with the needs of equipment connected by utility customers [2]. Supply interruptions and voltage dips are the two most common events affecting customers and have the largest financial impact on them. They disrupt commercial activities and manufacturing processes, resulting in decreased output and profitability [3]. Faults on transmission lines are a root cause of both interruptions and voltage dips [2].

The focus of this paper is on the classification of transmission line faults, according to the underlying cause, to meet the requirements for transmission systems control and operations.

The classification of faults plays a role in both the design and operational management of networks. Incorrect classification leads to uncertainty and error in developing ways to improve network reliability performance. The benefits of correct fault cause identification are a) reduced wasteful expenditure on inappropriate corrective measures, b) lower fault frequencies as the causes of faults are addressed, and c) automatic fault identification for immediate operational responses to faults.

Accurate identification of fault causes informs the design and parameter selection of new lines (insulator selection, tower design, footing resistance) and identifies poorly performing lines for implementing mitigating improvements. The design of new networks close to existing networks is often done on the basis that they will be exposed to similar conditions, and improvements can be made to reduce the effects of the prevalent causes of faults. For example, from the start of a project, bird guards might be adopted or clearances increased to minimize bird streamer faults.

In a typical scenario, once a fault has occurred, an operational crew is dispatched to patrol the line, identify the fault and conduct corrective work. Automatic classification can provide information for dispatchers to help teams to be appropriately equipped and prepared to look for characteristic fault signs. Where lines are not allowed to be returned to service without a complete line inspection to resolve uncertainty about the cause of a fault and concerns about equipment damage, early identification of the fault cause could allow a line to be restored more quickly [4].

The causes of faults are not always correctly identified as field services staff may give vague descriptions or lack knowledge about the fault mechanisms [5]. Thus, the aim of the present research is to establish an operational basis for classifying faults according to their underlying cause. The contributions of this work are: a) the development of waveform features to characterise faults across multiple voltage levels, at specific time intervals from fault initiation and according to the influence of the fault flashover mechanism; b) the ranking and selection of contextual and waveform features for identifying the causes of transmission line waveforms; c) the successful classification of transmission line faults by Single Nearest Neighbor classification, and d) showing that transmission line faults can be classified using either combined contextual and waveform features or only contextual features.

II. Literature Review

A. Waveform Characterization for Event Classification

Characterizing event waveforms (e.g. voltage dips) is used to reduce data, interpret and characterize events for analysis and management of power quality [6]. Methods include the ABC classification [7] and the South African NRS048-2 voltage dip characterization [8]. The requirement is to describe events with a limited number of parameters [9].

Characterization is also conducted for automatic classification of disturbances [10] with the aim being ‘to find common features that are likely related to specific underlying causes in power systems’ [11]. Various signal processing techniques, including root mean square (rms), Fourier and wavelet transforms [12], have been used to extract features and characterize events. They include finding the fundamental voltages and currents and harmonics, detecting transition points in waveforms, waveform segmentation and feature extraction [13].

Most studies base event classification on the disturbance type e.g. dips, transient and swells [13]. While this work is relevant for developing methods, its practical application is limited [10], since many studies with a shortage of data use synthetic data, leading to results with limited applicability to real-world scenarios [10].

Characterizing events according to their underlying causes is of more practical benefit, as described above, but limited work has been published. Classification of faults according to causes internal to the power system (e.g. transformer energizing, load changes and motor starting) has been explored by rms and Kalman filtering [14].

Characterization and analysis of external faults based on the voltage and current waveforms has been investigated for lightning, tree and animal contact, and cable faults [15]. Waveform characterization features were obtained from voltage and current waveforms recorded at distribution substations (12.47 kV) for 180 events, and included maximum zero sequence current and voltage, fault inception phase angle (FIPA), maximum change of current and voltage magnitudes (phase and neutral), and maximum arc voltage.

B. Classification of Power System Events

Pattern recognition is the science of information procedures for classifying, describing and labelling measurements [16]. A pattern recognition system includes stages of sensing, data pre-processing, feature extraction and classification [17]. Work towards developing pattern recognition techniques for recognition of power system events includes identifying the faulted phases e.g. single-phase-to-ground fault or phase-to-phase [18] and fault location [19]. Such studies use simulated and measured data from fault recorders on power systems. Less work has been done to identify the underlying causes of events [13]. One study uses the CN2 induction algorithm to determine rules to classify four causes of distribution network faults (lightning, tree, cable and animal) [15]. The CN2 rule induction algorithm induces an ordered list of classification rules from a set of classified observations [15]. This is an approach which is suitable for fault root cause identification where sufficient measurements with classifications are available.

Approaches to identifying animal-caused faults on distribution systems according to their root causes have included discrete wavelet transforms in combination with artificial immune systems [20], Bayesian networks [21], artificial neural [22] and fuzzy systems [23]. Artificial neural networks (ANN) and linear regression have been used to classify tree- and animal-caused faults based on distribution utility outage data [4]. Other methods applied to automatically diagnose the root cause of faults include support vector machines [24], expert systems for classifying events from measurements i.e. voltage step change, transformer energizing [14], as well as linear discriminant analysis [25]. Most studies have focused on distribution networks, with few aimed at root cause identification for transmission lines [24]. Such studies develop the theory and use of pattern recognition for identifying the causes of faults and power quality problems. This paper builds on the existing work by establishing a basis for classifying transmission system faults.

III. CHARACTERISING TRANSMISSION LINE FAULTS

A. Data Set

The transmission system of South Africa's electricity utility Eskom comprises over 28 000 km of lines operated at voltages of 132, 220, 275, 400 and 765 kV, of which the bulk are 400 kV and 275 kV lines. Records from 78 digital fault recorders (primarily SIMEAS-R and Siemens P513 devices) on the Eskom transmission network at 220kV, 275kV and 400kV over 13 years to 2008 were available for 2672 transmission line faults. Current and voltage waveforms are sampled at 2500 Hz.

Fault measurement records were checked to ensure they provided adequate pre-fault data and measurements from all voltage and current channels. The fault measurement records were matched using time, date and line on which the fault occurred to a database of 11573 faults developed by Minnaar *et al*. for the same transmission system [26]. This database of fault measurements is linked to the underlying fault cause as well as contextual information i.e. GIS data, line parameters and lightning density. This database addresses key concerns raised by Gu and Styvaktakis [13]: each characterized waveform is associated with a fault cause that makes this a suitable dataset for conducting feature selection and classification according to underlying causes; and the large data set, entirely based on measurements from an operational transmission system, addresses the limitations of many studies with too little data.

The original waveform data from digital fault recorders is stored in the IEEE C37.111 Common Format for Transient Data Exchange (COMTRADE). Each file represents a unique fault event measurement. The following data are available: sampling rate; start date and time; faulted phase; distance-to-fault; red, white and blue phase and neutral currents and voltages; number of samples; and timestamp data. The data was imported and stored in a Matlab structure. An array of structures, compiled with each individual measurement being a unique structure (file name is associated for identification), enables bulk signal processing of waveform data by repeating the same calculations inside a loop to obtain the desired characteristics. A symmetrical sequence component transformation was implemented in the Matlab Simulink environment, built around the discrete 3-phase sequence analysis block. The code to calculate waveform characteristics utilizes structure arrays in Matlab, so as to make possible the bulk signal processing necessary for 2672 waveforms. The Simulink model is then called from inside a 'for' loop to calculate the necessary parameters for each individual measurement, which in turn is stored inside a second structure array. The outputs of the Simulink model are discrete waveforms of magnitudes and phase angles for the positive, negative and zero sequence current and voltages. Sequence component currents and voltages are output in complex format and the rates of change for voltage and current sequence components are also exported. These are used to calculate a range of values for feature extraction and fault cause classification.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2015.2503478, IEEE Transactions on Power Delivery

B. Identifying the start and end of a fault

The features during a fault are required for characterization, making the identification of the start and end of a fault an important consideration. The three pre-fault (normal steady state operation), fault and interruption stages of measurement are illustrated for a single-phase-to-ground fault in Fig. 1. The start and end of the fault are identified using sequence components derived from the rms profiles.

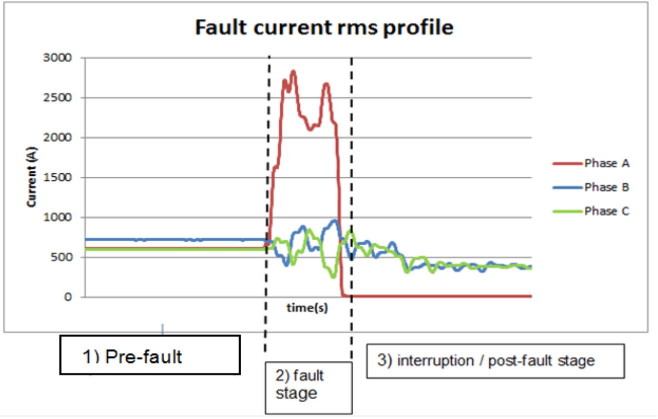

Fig. 1: Stages of a fault measurement – voltage and current

The beginning of the fault is identified by a detection index (dI), similar to that used for segmentation of rms voltage measurements [14], based on the difference between consecutive values of the positive sequence current:

$$dI = I_{1(n-1)} - I_{1(n)} \qquad (1)$$

A threshold set at 12000 kA/sec marks the start of a fault. In most records, measurement continues after the protection has operated, so post-fault values must be removed. A fault 'ends' when the 40ms moving average zero sequence current drops below 15% of the peak zero sequence current. Both the fault start and end thresholds were determined by inspection until the start point and endpoint of all the fault measurements in the dataset were captured.

C. Feature Extraction

The waveform features extracted in this research were chosen on the basis of their possible link to one of the root fault causes and their expected ability to identify a cause. These features were then tested for statistical significance linking them to a cause.

Characterization and analysis of external faults based on the voltage and current waveforms has previously been investigated for causes such as lightning, tree and animal contact and cable faults [15] and several features considered in that study were retained, albeit in a modified form. Maximum zero sequence currents and voltages are defined relative to pre-fault levels, enabling measurements at different voltage levels to be considered together, including relating the maximum values to pre-fault conditions. Factors such as fault level, pre-fault loading level, type and location of load, network configuration or capacitors being switched may influence the fault waveforms measured on the same line.

Several parameters from the measurement waveforms were calculated to test for these influences using the following strategies: 1) peak/maximum values relative to pre-fault values enable measurements from different voltage level networks to be considered together, 2) maximum changes of current were used to identify the influence of the fault flashover mechanism, and 3) the magnitude of characteristics at specific time intervals from the start of the fault were compared, characterizing the development of the fault.

The following fault features were extracted.

1) Faulted Phases

The relevance of the faulted-phases feature is based on the hypothesis that the physical flashover mechanism is a consequence of the underlying fault cause. The relationship is illustrated in Table 1, where the faulted phases (L) or ground (G) according to underlying cause are shown for the 220kV, 275kV and 400kV networks. More than 90% of all faults on the South African transmission network between 1995 and 2008 were single-phase-to-ground faults and pollution-caused faults are almost exclusively single-phase-ground-faults.

TABLE I
FAULTED PHASES ACCORDING TO UNDERLYING CAUSE

Faulted Phases	Bird streamer	Fire	Lightning	Other	Pollution	Total
L-G	1070	453	534	261	122	2440
L-L-G	15	33	66	8	1	123
L-L		25	1	2		28
L-L-L	30	4	32	4		70
L-L-L-G	4	3	3	1		11
Total	**1119**	**518**	**636**	**276**	**123**	**2672**

2) Maximum Change of Current ($\Delta I_{max0}, \Delta I_{max1}, \Delta I_{max2}$)

The maximum change of current during the initial transient stage, calculated using (2) as the maximum difference between consecutive samples after the fault has triggered, provides a picture of the dynamic state of the fault. The maximum change for each positive, negative and zero sequence component is treated as an individual feature.

$$\Delta I_{max} = max(I_{n+1} - I_n) \qquad (2)$$

This feature is chosen to test for a relation between the underlying cause and the rate of rise of fault current during the initial stages of a fault.

3) Maximum Sequence Voltage Ratio (V_{max2}, V_{max0})

The degree of unbalance [15] during a fault is calculated from the maximum negative and zero sequence voltages during a fault relative to their respective pre-fault values:

$$V_{rel\_\max\_i} = \frac{V_{fault\_\max\_i}}{V_{pre-fault}} \qquad (3)$$

Where i is either 0 or 2

4) Sequence Component Currents at ½ and 1 Cycle ($I_{0(0.5)}$, $I_{1(0.5)}$, $I_{2(0.5)}$, $I_{0(1)}$, $I_{1(1)}$, $I_{2(1)}$)

The values of the positive, negative and zero sequence component currents, illustrated in Fig. 2 for the zero sequence current component I_0, are measured at ½ cycle and one whole cycle after the fault trigger.

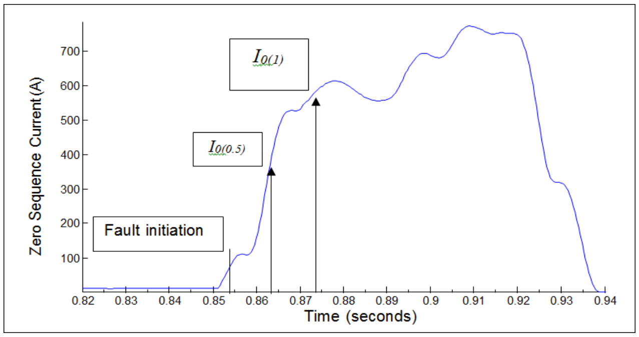

Fig. 2: I_0 at ½ and one cycle after fault trigger

5) *Maximum Sequence Current to Pre-Fault Current Ratio (I_{1max}, I_{0max})*

Peak values of sequence currents are calculated relative to the pre-fault current levels on the feeder.

$$I_{\max\_rel} = \frac{I_{peak\ post-fault}}{I_{peak\ pre-fault}} \qquad (4)$$

This gives an indication of the total state of change in current relative to the state of the network/load prior to fault occurrence. This relative value of fault current may provide more information than peak fault current, which is dominated by the network parameters. Maximum sequence current is therefore largely independent of network conditions or fault cause.

6) *Fault Resistance (*$R_{fault\_onecycle}$, $R_{fault\_twocycle}$*)*

The underlying mechanism by which a fault is formed and the medium along which fault current moves differs for each of the major fault causes. Bird streamer fault current flows via the liquid streamer, while fault currents due to fires are conducted via air and smoke particles. The resistivities of these mediums differ significantly. Similar to sequence currents, the fault resistance is calculated one and two cycles after the initiation of the fault (and denoted as $R_{fault\_onecycle}$ and $R_{fault\_twocycle}$) to account for the dynamic nature of faults. The line resistance from the point of measurement to the fault is based on the fault impedance values used for the protection settings for each line. The equations used to calculate fault resistance are based on the fault calculations according to Glover and Sarma [27]. The area of high impedance faults and the modelling of fault arcs, while worthy of investigation, are not dealt with in this paper. Table II shows the mean (μ) and standard deviation (σ) values of the Fault Resistance in ohms.

TABLE II
FAULT RESISTANCE

	220kV		275kV		400kV	
Characteristic	μ	σ	μ	σ	μ	σ
Rfault_onecycle	20.30	14.73	36.23	148.70	41.99	118.06
Rfault_twocycle	17.94	13.55	31.69	130.15	37.88	124.74

7) *Fault Inception Phase Angle (FIPA)*

Fault inception phase angle (FIPA) was considered in case the large dataset could give a clear indication of the relationship between fault types and fault peak. Its inclusion is based on statistical data presented by Barrera et al. that certain fault causes have fault inception angles near the peak of the waveform [15]. In particular, it is reported that cable faults and animal faults have average FIPA values of 93.7° and 99.3° respectively [15]. FIPA is calculated as the time of the trigger after the last zero-crossing prior to the fault. For multi-phase faults, FIPA is assumed to be the phase angle closest to the peak.

8) *Sequence Component Fault Current Time Constant*

The Sequence Component Fault Current Time Constant (τ0, τ1, τ2) is introduced as a waveform feature. It treats the fault waveform response in a similar manner to a first order linear time-invariant system. The time constant is calculated as the time taken from fault trigger to 0.63 of the difference between maximum fault current and pre-fault current for each sequence component, as illustrated in Fig. 3. The time constant for each sequence component current is an individual feature that may indicate the dynamic response of the transmission line/network to the fault type.

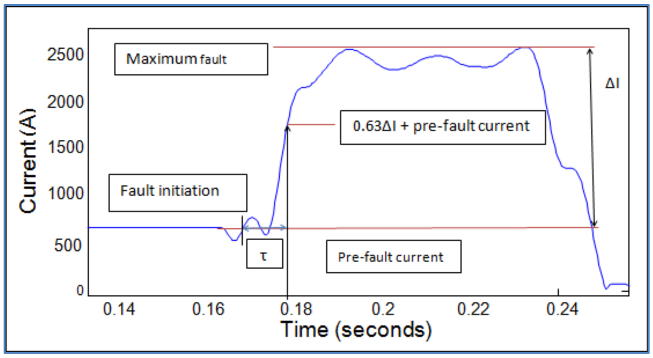

Fig. 3: Sequence Component Fault Current Time Constant

Table III shows the mean (μ) and standard deviation (σ) values of the Sequence Component Fault Current Time Constants in milliseconds [28]. These values are very similar across sequences phases; however they differ across system voltage levels. These results are indicative of most of the waveform characteristics extracted, illustrating the effect of the system rating on the magnitudes and changes in voltages and currents in response to faults. The voltage of the system on which a fault occurs should be included as a feature to analyze measurement data.

TABLE III
POSITIVE, NEGATIVE AND ZERO SEQUENCE FAULT CURRENT TIME CONSTANT STATISTICS

	220kV		275kV		400kV	
Characteristic	μ	σ	μ	σ	μ	σ
Positive current Time Constant (τ1)	12.81	1.48	26.85	70.01	22.72	71.41
Negative current Time Constant (τ2)	12.85	1.61	25.92	70.21	21.80	76.20
Zero current Time Constant (τ0)	13.46	7.14	22.31	46.26	20.27	63.65

D. *Statistical Significance of Waveform Features*

Analysing the waveform feature data by fault cause indicates that the fault cause influences the majority of them. The statistical significance (to the 0.05 level) of the causes (bird streamer, fire, lightning, pollution and other) on single-phase-to-ground faults occurring on 220kV, 275kV and 400kV networks, representing more than 90% of all faults, was tested by analysis of variance (ANOVA), for which results are given in Table IV ('yes' indicates statistically significant). The differences across voltage levels for all the waveform features are statistically significant, with the exception of FIPA, indicating that fault causes may be differentiated by a combination of these features.

TABLE IV
STATISTICAL SIGNIFICANCE OF CAUSES INFLUENCING WAVEFORM FEATURES

FEATURE	400kV	275 kV	220kV
ΔI_{max1}, ΔI_{max2}, Rfault_onecycle, Rfault_twocycle	yes	yes	yes
ΔI_{max0}, $I_{1(0.5)}$, $I_{1(1)}$, $I_{0(1)}$, I_{1max}, I_{0max}	yes	no	yes
V_{max2}, V_{max0}	yes	no	no
$I_{2(0.5)}$, Pos. current Time Constant (τ1)	yes	yes	no
$I_{2(1)}$, $I_{0(0.5)}$, I_{2max}	no	no	yes
FIPA	no	no	no
Negative current Time Constant (τ2), Zero current Time Constant (τ0)	no	yes	no

IV. Fault-Root Identification Based on Pattern Recognition

This section explores the classification of transmission line faults according to underlying cause using pattern recognition techniques. The individual relevance of features is determined with respect to the four major causes identified and classification is treated as a multiclass problem. Features are ranked according to their relevance in separating fault causes and classifiers are built by nested subsets. The full feature set consists of 7 contextual features (environmental, climatic and diurnal) and 21 waveform features (shown in Table IV, including system voltage).

A key finding is that the fault frequencies have statistically significant differences with respect to time-of-day, season, and climate as represented by rainfall area in South Africa. The differences are not uniform across the voltage levels; hence voltage level is also a feature. The contextual feature set also introduces other features considered relevant. Thus the contextual features describing fault occurrence are hour of day, month of year, rainfall area, voltage level, line GIS number, Eskom transmission grid region and the lightning ground flash density [26].

In this study, feature selection was used first to improve understanding and interpretability of the data and secondly to identify features for building good classifiers for transmission line fault causes. Feature ranking and classification was considered for three scenarios, using: a) only the contextual feature set, because earlier analysis had shown statistically significant differences in fault frequencies by time of day, climate and season; b) only the waveform feature set; and c) combined waveform and contextual feature sets. Feature selection and classification was implemented in the Matlab toolbox PRTOOLS [29].

A. Feature Ranking

Feature ranking according to the F-statistic derived with ANOVA provides a measure of variance due to each feature and a basis for building classifiers.

TABLE V
FEATURES RANKED BY F-STATISTIC

Feature	Type of Feature	F-statistic	Overall Rank
Hour	Contextual	90.90	1
Region	Contextual	33.09	2
Month	Contextual	32.16	3
Nominal Voltage	Contextual	29.22	4
Average ground flash density (Ng)	Contextual	25.45	5
$I_{2(0.5)}$	Waveform	16.06	6
τ1	Waveform	11.98	7
Faulted Phases	Waveform	11.58	8
V_{max2}	Waveform	10.44	9
τ2	Waveform	9.97	10

From a full ranking [28], Table V lists the top ten features and gives a clear picture of the relevance of each for distinguishing faults according to cause. The waveform features with the highest F-statistic scores are the maximum negative sequence current (half cycle) and the positive sequence time constant labelled $I_{2(0.5)}$ and τ1 respectively. Both provide a measure of the dynamic response to an event on a transmission line, relating to the rate at which sequence currents rise to peak fault current. Table V indicates that three individual contextual features related to time of occurrence and geographic location of a fault are highly relevant to identifying its underlying cause. This result is consistent with the success achieved using only contextual features for classifying animal- and tree-caused faults on distribution systems [24]. The ranking also gives insight into the relative strength of using the contextual and waveform features to identify fault causes.

B. Classification by Nested Subsets

The classification problem for transmission line faults is defined here as a multiclass problem with five classes. A multiclass classifier is a function F:X →***Y*** which maps an instance **x** into a label **F(x)** [30].

There are two common approaches to generating ***F***. The first is to construct it through a combination of binary classes, e.g. logistic regression or support vector machines; and the second (used in this study) is to generate ***F*** directly e.g. naïve Bayes or nearest neighbor (NN) algorithms. The 1-NN classifier is proposed for identifying the underlying cause of transmission line faults.

The 1-NN rule is a suitable benchmark for other classifiers as it requires no user-specified parameters, making it implementation independent, and provides reasonable classification performance in most applications [31]. Building a series of 1-NN classifiers determines the underlying cause of faults. An advantage of 1-NN classification is its conceptual simplicity and ease of implementation [32].

The 1-NN classifiers comprise five classes, one for each major fault cause: birds, fire, lightning, pollution, and other. The classifier finds the nearest point in the training set to the unclassified point and assigns it to the corresponding label. Feature selection for the initial classification is done by building nested subsets of features of increasing size. Classifier building starts with a subset of one feature (the highest F-statistic) and features of decreasing F-statistic are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2015.2503478, IEEE Transactions on Power Delivery

progressively added. For example, using Table V, the first subset comprises the feature ranked 1 and the second subset the features ranked 1 and 2.

A classifier is trained using a training set of two thirds of the data and performance is evaluated using the test set of the remaining third of the data. The entire data set is randomly split 30 times into the training and test sets, and the trained classifier evaluated against each test set to reduce variance from the classification results. Evaluating the classifier in this manner provides enough test data for the faults causes that occur rarely i.e. other- and pollution-caused faults.

Classification is conducted according to the three scenarios: using only contextual features, only waveform features, and the combined feature set. Two combining rules are implemented for classification using all the features. The first (Rule 1) combines waveform and contextual feature sets by relevance using the overall ranking as indicated in Table IV; the second (Rule 2) combines the waveform and contextual feature sets by adding (in order of decreasing relevance) a feature from each set, starting with the contextual set. Once all the contextual features are added, the remaining waveform features are added to form additional subsets.

C. Assessing Classifier Performance

Many classifier's performance measures are based on the confusion matrix. A 2x2 confusion matrix for a 2-class classifier (Yes/No) and a test dataset is shown in Table VI [33].

TABLE VI:
CONFUSION MATRIX

Class	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The most common performance measure calculated from this matrix is Accuracy (Acc), which is the proportion of the total number of predictions that were correct:

$$Acc = \frac{TN+PN}{TP+FN+FP+TN} \qquad (5)$$

An acceptable Accuracy rate for practical classification may be set at 80%. However, Accuracy is an insufficient performance measure in applications with imbalanced data sets (i.e. one class represents only a small portion of the total data) because a classifier ignoring the presence of the minority class will show good performance. The fault performance of the transmission system is unbalanced across the four major fault causes.

Alternative measures for classifying unbalanced datasets include Precision (the percentage of correct positive predictions), Recall (the percentage of true positive cases correctly identified), and the F-measure [34], which is the harmonic mean of Precision and Recall given by

$$F-measure = \frac{2\times Precision\times Recall}{Precision+Recall} \qquad (6)$$

(Note: the F-statistic used in ANOVA and the F-measure are two independent measures defined and used in different ways). In this study, the measures used to assess classifier performance are overall classification Accuracy and the F-measure for each fault cause.

D. Overall Classification Accuracy

The overall classification accuracy rates indicate that classification accuracy up to 90% is achieved when using only the five highest ranked contextual features.

Fig. 4 illustrates overall classification accuracy for contextual, waveform and all features. The features are added according to the respective decreasing rankings shown in Table IV for waveform and contextual features and then according to the combining rules defined. For example, for ACC_All_rule1 all 21 features of Table IV are added according to Rule 1.

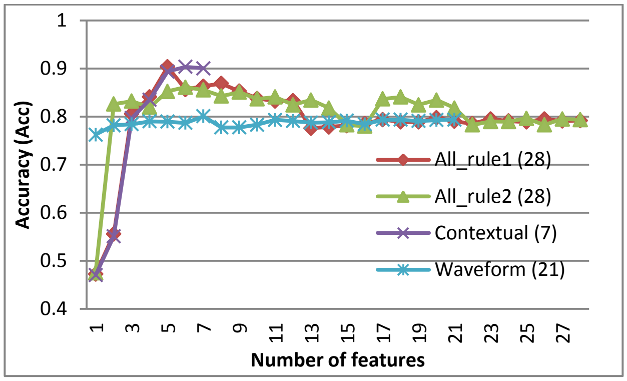

Fig. 4: Classification Accuracy

Reasonably good classification performance can be achieved using only waveform features; the best accuracy achieved is 0.801 using the seven highest ranked features. It does not match the performance achieved using only contextual features or combined contextual and measurement features.

E. Classification Performance using Waveform Features

Figure 5 illustrates the classification success by F-measure for each of the major fault causes, adding features according to decreasing F-statistic. Scores above 0.75 are achieved using only the two highest ranked features for bird, fire and lightning caused faults. The F-measure for pollution-caused faults is generally lower than the first three classes, but it needs to be considered that pollution is a highly imbalanced set (pollution faults represent only a small portion of the total data). The performance of the 'other' class of faults is the poorest for most feature sets and depends on an appropriate feature set being selected to achieve reasonable performance.

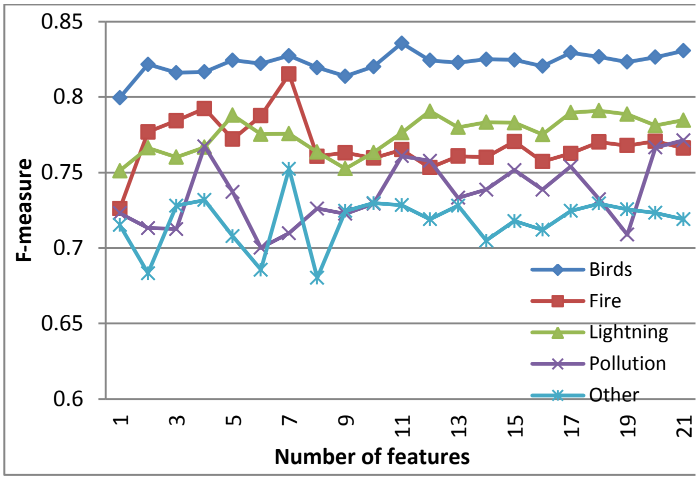

Fig. 5: F-measure using waveform features

F. *Classification Performance using Contextual Features*

The classification performance achieved using contextual features is significantly better once five or more features are used (with features added by decreasing F-statistic) to build the 1-NN classifier. F-measure scores above 0.8 are achieved for all classes of faults with bird streamers having the highest score of 0.917, as illustrated in Fig. 6.

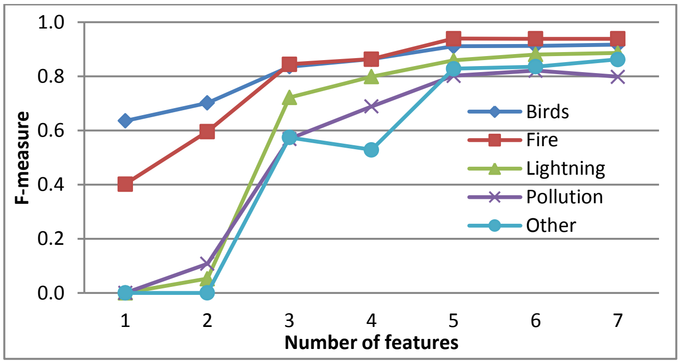

Fig. 6: F-measure using contextual features

G. *Classification Performance using Combined Features*

Combining the waveform and contextual features does not improve the classification over that achieved by using only contextual features. Fig. 7 shows the F-measure performance achieved using combining-rule 2.

The best classification accuracy of 0.86 is achieved using six features (the top three ranked from the waveform and contextual sets): Hour, Negative sequence current (half cycle after fault initiation), Region, Positive sequence fault current time constant, Month and Faulted Phases. The best performance achieved with the 1-NN classifier is compared with the classification performance of several common classifiers with feature selection conducted by sequential forward (SFS) and sequential backwards (SBS) wrapper selection based on best performance for a particular classifier. The classifiers used for the comparison are radial basis neural network, decision tree and naïve Bayes classifiers.

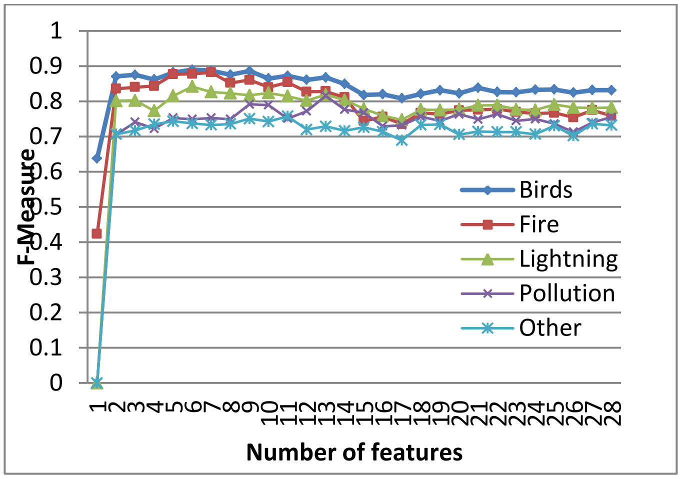

Fig. 7: F-measure using waveform and contextual features combined by rule 2

TABLE VII:
CLASSIFICATION PERFORMANCE

Classifier	Decision tree	Decision tree	Neural Network	Neural Network	Naïve Bayes classifier	Naïve Bayes classifier	1-NN (combining rule 2)
Selection	SFS	SBS	SFS	SBS	SFS	SBS	Nested subsets
ACC	0.74	0.73	0.74	0.74	0.74	0.74	0.86
F-measure Birds	0.78	0.77	0.78	0.78	0.80	0.78	0.89
F-measure Fire	0.82	0.81	0.81	0.81	0.78	0.80	0.88
F-measure Lightning	0.67	0.65	0.67	0.73	0.68	0.68	0.84
F-measure Pollution	0.31	0.00	0.30	0.28	0.29	0.34	0.75
F-measure Other	0.17	0.20	0.16	0.15	0.14	0.15	0.74

The comparative analysis of classification results shown in Table VII demonstrates that the best performance achieved with the 1-NN classifier is between 12-14% higher than achieved with any of the other classifiers using either forward or backward selection. The 1-NN classifier is also shown to have considerably better classification performance when considering the minority causes (Pollution and Other causes).

V. Implications of Results

Classifiers developed using only the contextual features demonstrate the highest level of balanced classification performance (F-measure=90%). The relevance of this is that in instances where no physical evidence of the fault is present and operators assign fault causes based on contextual information (e.g. a fault during a thunderstorm may not be due to lightning but is often classified as such [15]), the practice overlaps directly with the best classification performance. However, the lower classification accuracies achieved with waveform features may be relevant when they point to results differing from those indicated by the contextual features.

Using only waveform features to build a classifier produces reasonable success levels, even with only the single most

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2015.2503478, IEEE Transactions on Power Delivery

relevant feature. While this is an important finding for identifying and using fault waveform features for fault identification, the classification performance achieved may be inadequate for practical applications. An acceptable level of accuracy (80%) is achieved with only contextual features and by combining contextual and waveform features.

The concern with classifying faults using only contextual features is that actual measurements or observation of the event are not considered. In practice, physical observations, e.g. flashover markings on towers, are important for confirming fault cause.

The results of automatically classifying faults according to cause demonstrate the suitability of pattern recognition techniques (particularly 1-NN classifiers) for practical applications. Classifying the causes of faults within the operational timeframe applicable in a control center has the potential to improve transmission system reliability. The relevant timeframe has been identified as less than five minutes [35]. This work lays the foundation for implementing automatic classification of transmission line faults within the operational requirements of a transmission system, following the structure shown in Fig. 8. The following are the requirements for implementing real-time classification:

a) Knowledge of the power system network, its parameters, external geography and climate play a key role in identifying fault causes. These contextual features should be tabulated in a dataset that can be linked to fault measurement waveform data.
b) Appropriate feature extraction from waveform data is critical. This may be done either on the fault waveform recorder (or its associated software) or in a centralized database. For faster classification, it is preferable to conduct post-processing and feature extraction on the fault waveform recorder. Where legacy devices are in use, without the requisite software or processing power to conduct post-processing and feature extraction, it is recommended that waveforms should be retrieved to a central database for centralized feature extraction.
c) Waveform and contextual features should be combined in a single database forming the platform for classification.

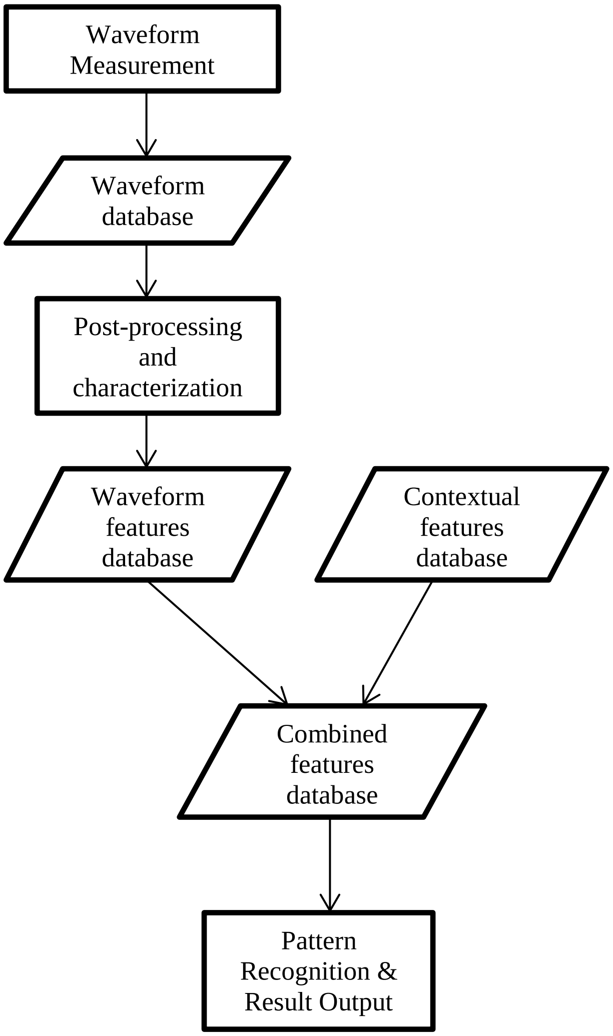
Waveform Measurement
Waveform database
Post-processing and characterization
Waveform features database
Contextual features database
Combined features database
Pattern Recognition & Result Output

Fig. 8: Structure for operational fault classification

VI. CONCLUSIONS

It has been shown that both contextual and waveform features can be identified which have a varying degree of relevance to performing classification of events. The accuracy of classification using these features has been demonstrated by building classifiers using nested subsets. Best classification is achieved using only contextual features; however classification in this manner does not make use of any measurements. Taking waveform features based on measurements into account, classification accuracy of 86% is achieved using the following set of six contextual and waveform features: Hour, Negative sequence current (half cycle after fault initiation), Region, Positive sequence fault current time constant, Month and Faulted Phases.

The highest classification accuracy achieved using only waveform features is 80%. This indicates that faults can be classified for underlying cause using only waveform features with a reasonable level of success. However this performance is inferior to classification accuracies achieved when using waveform features in combination with contextual features.

This research shows that the underlying cause of transmission line faults can be classified automatically. The results achieved indicate that the 1-NN classifier is a suitable classifier for identifying the causes of transmission line faults and achieves superior classification in comparison to other classifiers. This work lays a foundation for operational classification of transmission line faults and the key requirements and structure of a fault classification system have been developed.

ACKNOWLEDGEMENTS

The contribution of fault and measurement data by Eskom Holdings is acknowledged.

REFERENCES

[1] L. Alvehag, L. Soder, "A reliability model for distribution systems incorporating seasonal variations in severe weather" *IEEE Trans. Power Delivery,* vol. 26, no. 2, pp. 910-919, Apr. 2011.

[2] Cigre, *Economic Framework for Power Quality, JWG Cigre-CIRED C4.107,* 2011

[3] A. Chowdhury, D. Koval , "Development of transmission reliability performance benchmarks" *IEEE Trans. Industrial Applications,* vol. 36, no. 3, pp. 899-903, Jun. 2000.

[4] L. Xu, M. Chow, "A classification approach for power distribution systems fault cause identification", *IEEE Trans. Power Systems,* vol. 21, no.1, pp. 53-60, Feb. 2006.

[5] H. Vosloo, "The need for and contents of a life cycle management plan for Eskom transmission line servitudes", M.Sc Thesis, Dept. Geography, Univ. Johannesburg, Gauteng, South Africa, 2005.

[6] *Electromagnetic compatibility (EMC) - Part 4-30: Testing and measurement techniques - Power quality measurement methods,* IEC61000-4-30, 2008.

[7] M. Bollen, "Algorithms for characterizing measured three-phase unbalanced voltage dips", *IEEE Trans. Power Delivery,* vol. 18, no. 3, pp. 937-944, 2003.

[8] *Electric Supply - Quality of Supply Part 2: voltage characteristics,compatability levels, limits and assessment methods, NRS 048-2:2007,2007*

[9] R. Koch, A. Botha, P. Johnson, R. McCurrach , R. Ragoonanthun, "Developments in voltage dip (sag) characterisation and the application to compatibility engineering of utility networks and industrial plants", *Proceedings of the 3rd Southern African Power Quality Conference.* Livingston, Zambia Oct. 2001.

[10] M. Bollen, I. Gu, E. Styvaktakis, "Classification of Underlying Causes of Power Quality Disturbances: Deterministic versus Statistical Methods" *EURASIP Journal on Advances in Signal Processsing,* vol. 1, pp. 1-17. Feb. 2007.

[11] M. Bollen, I. Gu, S. Santoso, M. Mcgranaghan, P. Crossley, M. Ribeiro, "Bridging the gap between signal and power", *IEEE Signal Processing Magazine,* vol. 26, no. 4, pp. 12-31. Jul. 2009.

[12] R. Fernandez, H. Rojas, "An overview of wavelet transform in application in power systems", *Proceedings of the 14th Power Systems Computation Conference,* Sevilla, Spain, Jun. 2002.

[13] I. Gu, E. Styvaktakis, "Bridge the gap: Signal processing for power quality applications", *Electric Power Systems Research, vol. 66, no. 1,* pp. 83-96, Jul. 2003

[14] E. Styvaktakis, M. Bollen, I. Gu, "Expert system for classification and analysis of power system events", *IEEE Trans. Power Delivery, vol. 17, no.2,* pp. 423-428, Apr. 2002

[15] V. Barrera , J. Melendez, S. Kulkharni, S. Santoso, "Feature analysis and automatic classification of short circuit faults resulting from external causes", *European Transactions on Electrical Power,* vol.23, no.4, pp. 510-525, Jan. 2012.

[16] P. Jonker, R. Duin, D. De Ridder, "Pattern recognition for metal defect detection" *Steel Grips, vol. 1,* no.1, pp. 20-23. 2003

[17] R Duin, F. Roli,D. De Ridder, "A note on core research issues for statistical pattern recognition", *Pattern Recognition Letters,* vol. 23, no.1, pp. 493-499, Feb.2002

[18] V. Ziolowski, I. da Silva, R. Flauzino, "Automatic identification of faults in power systems using control technique" *Proceedings of thee 16th International Conference on Control Applications.* Singapore, 2007.

[19] J. Mora-Florez, J. Cormane-Angarita, G. Ordonez-Plata, "K-means algorithm for power quality applications", *Electric Power Systems Research,* vol. 79, no. 5 , pp. 714-721, 2009.

[20] L. Xu, M. Chow, "Distribution fault diagnosis using a hybrid algorithm of fuzzy classification and artificial immune system", *Proceedings of IEEE PES General Meeting - Conversion and Delivery of Electrical Energy.* Raleigh, 2008

[21] R. Teive, J. Coelho, P. Charles, T. Lange, L. Cimino, "A bayesian network approach to fault diagnosis and prognosis in power transmission systems", *Proceedings of the 16th International Conference on Intelligent System Application to Power System (ISAP).* Crete, 2011

[22] M. Chow, S. Yee, L. Taylor, "Recognizing animal-caused faults in power systems using artificial neural networks" *IEEE Transactions on Power Delivery,* vol. 8, no. 3, pp. 1268-1273, Jul. 1993

[23] S. Meher, A. Pradhan, "Fuzzy classifiers for power quality event analysis" *Electric Power Systems Research,* vol. 80,no.1 , pp. 71-76, 2010

[24] B. Ravikumar, D. Thukaram, D., H. Khincha, "Application of support vector machines for fault diagnosis in power transmission system", *IET, Generation, Transmission and Distribution, vol. 2. no.1*, pp. 119-130, Jan. 2008

[25] Y. Cai, M. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in smart grids", *IEEE PES General Meeting.* Calgary, Jul. 2009

[26] U. Minnaar, C. Gaunt, F. Nicolls, "Characterisation of power system faults on South African transmission power lines" *Electric Power Systems Research,* vol. 88, nol.1, pp. 25-32,Jul. 2012

[27] J.D. Glover, M. Sarma , *Power System Analysis and Design.* Pacific Grove, Brooks/Cole, 2002.

[28] U.J. Minnaar, "The Characterisation and Automatic Classification of Transmission line faults according to Underlying Cause" Ph.D. Dissertation, Elec. Eng., UCT, Cape Town, South Africa, 2014

[29] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, D. Tax, *PRTools 4, A Matlab Toolbox for Pattern Recognition.* Delft University of Technology,Delft, 2000.

[30] R. Duda, P. Hart, D. Stork, *Pattern Classification.* Wiley-Interscience, New York, 2000.

[31] A. Jain, R. Duin , J. Mao, "Statistical pattern recognition: a review", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37,Jan. 2000.

[32] V. Garcia, R. Mollineda, J. Sanchez, "On the k-nn performance in a challenging scenario of imbalance and overlapping", *Pattern Analysis and Applications,* Vol. 11, pp. 269-280,Sept. 2008

[33] M. Kubat, R. Holte, S.Matwin, "Machine learning for the detection of oil spills in radar images" *Machine Learning,* vol. 30, no. 1, pp. 195-215, Feb. 1998

[34] Y. Nan, K. Chai, W. Lee, H.Chai, H, "Optimizing f-measure: a tale of two approaches" *29th International Conference on Machine Learning.* Edinburgh, Jun. 2012

[35] J. Bekker, P. Keller, "Enhancement of an Expert System Philosophy for Automatic Fault Analysis", *11th Annual Georgia Tech Fault and Disturbance Analysis Conference*, Atlanta, 2002, Available: http://truc.org/media/1292/enhancement-of-an-expert-system-philosophy-for-automatic-fault-analysis.pdf

Ulrich Minnaar received the BSc (Eng) in 2001, MSc (Eng) in 2006 and PhD in 2014. **Trevor Gaunt** is an Emeritus Professor in the Department of Electrical Engineering at the University of Cape Town. **Fred Nicolls** is an Associate Professor in the Department of Electrical Engineering at the University of Cape Town.