The effectiveness of different sampling rates in vegetation high-impedance fault classification

The effectiveness of different sampling rates in vegetation high-impedance fault classification

Douglas P. S. Gomes, Cagil Ozansoy,
Anwaar Ulhaq, José Carlos de Melo Vieira Júnior

Abstract

This paper investigates the alarming fire igniting scenario of High-Impedance Faults (HIF) resulting from the contact of vegetation with the power lines. Our findings are based on a set of experiments performed on a data set of real staged fault-signals sampled by two distinct channels with different band-pass filters. Representations from these two sampling methods are extracted by different signal processing methods and ranked by their discriminative potential. Experimental results obtained by our proposed methodology show the effectiveness of wide band signals sampled at higher frequencies. Their features result in higher separability potential and are more effective at discriminating fault occurrences than ones from the low-frequency channel. As the approach of employing high-frequency signals in such task may be faced with skepticism, the paper also discusses the possible concerns and feasibility of using higher sampling rate technologies for HIF fault classification.

Keywords

Fire mitigation, high-frequency, high-impedance faults, vegetation faults.

1 Introduction

In the occurrence of an electric fault, the resulting fault current amplitude might not exceed the threshold that triggers protection devices. Such events are specifically treated in the related literature as High Impedance Faults (HIFs). In the electrical distribution systems context, these distinct events represent an enduring and vastly discussed subject [1].

Nevertheless, despite their lasting history, there isn't a consensus regarding a definite solution or methodology for detecting such events. In fact, a global solution for HIFs may be impractical given the numerous factors that influence the phenomena behavior. Examples are network grounding type, voltage level, fault impedance value, signal sampling parameters, and more importantly, the fault contact surface [2, 3, 4, 5]. Recent works [6, 7, 8] strongly argue that a possible solution may be achievable if HIFs receive special treatment regarding their specific types. In [6], the present authors propose a methodology for fault detection inspired in existing methods but solely addressing vegetation HIFs. A model for tree-related HIFs is proposed in [7], with the premise that there is no single model that represents the expected behaviors of such faults in the literature. A further paper by the same authors [8] uses the same premise to propose a tree-related HIF location methodology. By focusing on a specific type of fault and challenging the limitations of low-frequency sampling, such works indicate that the aforementioned definition may be too limited for the set of intricate problems related to HIFs.

In this manner, the present paper focuses its efforts on investigating a specific type of fault, labeled here as vegetation HIFs. This definition, used throughout this document, relates to all scenarios where powerlines come in contact with vegetation, resulting in single-digit ampere fault currents. Powerlines breaking and falling to vegetation at ground level, vegetation brought by heavy winds bridging two phase conductors, or tall trees reaching powerlines are examples of such scenarios. Similar to general HIFs, these faults do not represent a great risk regarding equipment stress. Their importance is rather related to their ability to ignite fires. Once in contact with vegetation, powerlines pose a significant fire risk, even if the resulting current has a single-digit ampere value in amplitude [9].

Although this is of special concern for certain weather and flora conditions such as the one often encountered in the state of Victoria, Australia, countries such as United States, Spain, and Brazil have been associated with fires created by power distribution lines [9, 10, 11]. Relevant weather conditions are high temperatures, low humidity, rain frequency, the proportion of dry vegetation, and wind strength [12]. Australia's history with bushfires is so dense that, after dramatic events in 2009, a fire risk mitigation program called Powerline Bushfire Safety Program (PBSP) was created by the Victorian government [13]. It resulted in a variety of network augmentations and research projects, such as the 'Vegetation ignition testing', which performed hundreds of staged vegetation faults resulting in the data set further analyzed in this paper.

The challenges of detecting vegetation HIFs are often connected to the limitations of existing hardware, the ambiguity of measurements throughout distribution feeders, and the fact that fault currents cannot be distinguished from a simple increase in customer load. These constraints often guide the most relevant HIF detection methods proposed in the literature [14]. The obtained processed signals are generally resulted from simulations or staged faults in laboratories [14]. They also do not contemplate high-frequency (>10 kHz) components since their sampling rate is chosen using existing hardware sampling capabilities [6]. Nevertheless, a different approach was taken by the authors in [6] by proposing a vegetation HIF detection methodology that relied on features calculated from High-Frequency (HF) sampled signals. The investigation described in the present paper follows such context by presenting a comparison between the discriminative potential of features calculated from the previously discussed HF sampling, in contrast to the traditional low-frequency approaches. The signals used in the analysis were sampled from staged vegetation HIFs, made in a real loaded 22 kV network, with high resolution, and wide-band recordings. The derived conclusions were strongly connected to the network neutral connection type, which in this case is given by a compensated (resonant) grounding scheme. This means that the system operates with an isolated or a non-solidly earthed neutral connection. Some Nordic countries [3] and Australia's distribution systems can be cited as examples of such use. Nevertheless, it must be stated for clarification purposes, that this paper does not propose a HIF detection method like many in the related literature. It rather intends to add value to researchers and industry when considering the development of high-frequency sampling HIF detection methods.

The diversion (high-frequency investigation) from the standard (low-frequency) fault detection methodology is often followed by understandable and strong skepticism given by its practicality and feasibility. Therefore, this paper also presents relevant related points for a justifiable concern and attention to these non-standard, evolving sampling technologies. These arguments are given in the discussion section, which precedes the description of the dataset characteristics, the disclosure of the utilized methodology, and its subsequent results, each one in their following respective sections.

2 Vegetation faults dataset

In order to clarify methodology decisions and problem background, the following sections discuss the origins, characteristics, and preparation of the analyzed dataset.

2.1 Origins and data collection

The state of Victoria in Australia is no stranger to bushfires associated with electric distribution systems. One drastic example is the fires of "Black Saturday" in 2009. A series of fifteen fires that collectively burnt over 270 000 ha, caused more than 150 fatalities, and destroyed more than 1800 homes [15]. Six of these major fires were attributed to faulty electric assets.

These events led to the creation of the Powerline Bushfire Safety Program (PBSP), an initiative to reduce the risk of bushfires caused by distribution systems. Such goal was to be achieved by carrying out legislative measures, network augmentations, and funding of research projects related to new fault detection technologies. The 'Ignition Conduction Ignition Testing' [6], a program responsible for sampling local vegetation species and using them to stage vegetation faults on a real 22-kV feeder, was one of these projects. By the end of it, more than a thousand of tests were performed with different species and fault configurations.

Despite much research, nine years after the Black Saturday, fires ripped through country Victoria between March 17-18, 2018. These fires, now known as St. Patrick day fires, destroyed homes and killed many livestock. At least two of these fires were powerpole/powerline related, attesting for the inadequacy of overcurrent methods at protecting against HIFs hazards.

2.2 Fault tests

The team responsible for the tests made numerous valuable and a few questionable decisions in the test arrangements and project design. Between the beneficial, the decision of sampling the fault signals with high resolution, low-noise, in wide-band recordings was crucial. Likewise, the idea of contemplating different fault configurations labeled as 'Branch touching wires' (phase to earth), 'Branch across wires' (phase to phase), and 'Wire into vegetation' (phase to earth) faults. In fact, the decision to include phase-to-phase tests represent one of the most valuable aspects of the resulted dataset. HIFs are disturbances often modeled by a scenario where a conductor breaks and falls to a surface such as asphalt or gravel. These are phase to earth faults, that despite their relevance, have different arcing characteristics from a vegetation phase-to-phase fault involving only phase currents. Moreover, the decision to set the majority (99%+) of the tests' fault current thresholds between 0.5 and 4 A was key to the program findings and insights. The thresholds were set with the help of high-voltage resistors to replicate the nature of HIFs in real life. They helped prove that such single-digit fault currents can still ignite fires in contact with vegetation, and that such fire risk could be significantly reduced if the faults were detected in five seconds or less [6].

The former, in particular, was reinforced by having a well-known commercial protection relay with an embedded HIF detection function that did not detect any of the staged faults [6], corroborating the inadequacy claim from current technologies at detecting such faults. The outcomes of such experiments can not only aid fault detection research but also as to set guidelines for future detection technologies. Additional findings include: assessment of the fire ignition probability at different current values, study of the ignition phases in vegetation conduction, and the effect of chemical composition and species' type regarding their likelihood to ignite fires. For example, the species found to be the most dangerous regarding fire risk were Willows, and the safest ones to be Peppercorns.

It is important to specifically note that the mentioned current thresholds represent a massive deviation to the ones found in standard HIF detection research literature. Since the definition usually used for HIFs embraces all the faults that overcurrent devices cannot detect, the assumed current fault values are often much higher [14, 16, 17]. The fact such fault currents can ignite fires, despite having small amplitudes, represents an important problem to be considered. This can be highlighted as the most relevant arguments for the need for higher sensitivity protection devices in vegetation areas with high risk of fire ignition, and the treatment of HIFs by specific scenarios. In the state of Victoria, which has compensated networks (known by their smaller earth-fault currents), protection device sensibility ranges between 5 to 10 A.

The not so valuable decisions, however, was mainly related to the test methodology design. The choice of placing a vegetation sample between the conductors prior to their energization can be firstly cited as the most important. It resulted in instant conduction after the energization of the conductors, excluding the possibility of pre-fault signal analysis. Nevertheless, a strategy adopted to counter such relevant drawback, discussed in the next section, allowed the creation of non-fault observations and further comparative analysis. In like manner, another constricting decision was to stage the faults in a dedicated feeder, constructed purely for the sake of tests. This implied that, although connected to a loaded network, the sampled signals did not include the load current contribution. Such drawback drove the present analysis to discuss the information content of voltages, rather than current signals. In this manner, it can certainly be considered as a shortcoming from the dataset but, when placed in the context of the practical high sampling rate hardware (expanded in discussions), it is realized that the voltage signals is indeed the domain where such investigations can be useful.

2.3 Sampling

Throughout the fault experiments, both fault current and voltage signals were sampled in two different channels. That is, two electrical quantities were sampled via two channels at the same time. Such redundancy was chosen due to the intention to characterize the signals in a wideband with low-noise. These channels were unambiguously labeled as Low-Frequency (LF) and HF channels due to their bandwidth. The first had a low-pass filter with 50 kHz corner frequency, sampling signals at 100 kSa/s in a continuous sampling mode. The HF channel was connected to a high-pass filter, sampling frequencies greater than 10 kHz at 2 MSa/s in a sweep sampling mode. By sweep sampling, the authors mean that signals were only sampled by a brief duration throughout a certain period. In the tests, the sampling period was 1 second, and the sweeps had 20 ms of duration (one power cycle at 50 Hz). As a noticeable example, such sampling duration is enough to comprise at least 200 full cycles in a single sweep at the lowest frequency in the HF channel's bandwidth (10 kHz).

2.3.1 Fault observations

The fault testing program reported 1038 different staged faults. The present analysis, however, could not make use of all reported recordings. Tests that had missing, invalid or corrupted data, high intermittency or no current conduction, for example, were excluded. Such cleansing led to a final number of 768 distinct experiments wherein each test had only one sweep extracted for analysis. The procedure of extracting the analyzed sweep was one of the pre-processing practices made on the dataset. It basically selected the first sweep to be sampled after the RMS current value met a chosen threshold. In this paper, in order to clearly illustrate the powerful information content of the HF signals, a pessimistic threshold was used, being the value of 0.1 A.

The discussed types of faults were present in the extracted sweeps (observations) with phase-to-earth faults attesting for 68.75% of the fault type labels, while phase-to-phase represented the remaining 31.25%. This can be highlighted as of additional value since the results soon to be presented were consistent with both types of faults. The main perceived difference was that phase-to-phase faults had higher RMS voltage values which resulted in higher currents, making them easier to discriminate than phase-to-earth faults. Yet, despite frequent, such observation could not be used as a reliable fault predictor since it did not present relevant discriminative power between classes.

2.3.2 Non-fault observations

As discussed, constraints in the experiment design made it impossible to obtain pre-fault data. However, the project's team decided to make recordings of the network's electric signals operating at a steady state throughout the tests days. The goal was to gather data to help characterize the standard background noise of the studied feeder. Indeed, a useful idea that resulted in a large number of sweeps from both channels which were then used as Non-fault observations.

Regarding the present investigation as a comparative analysis, a decision was made to have the same number of observations of both signal states. These sweeps were sampled at random between all tests days that had such recordings, adding up the number of total observations to 1536 examples, 768 from each Non-fault and Fault states.

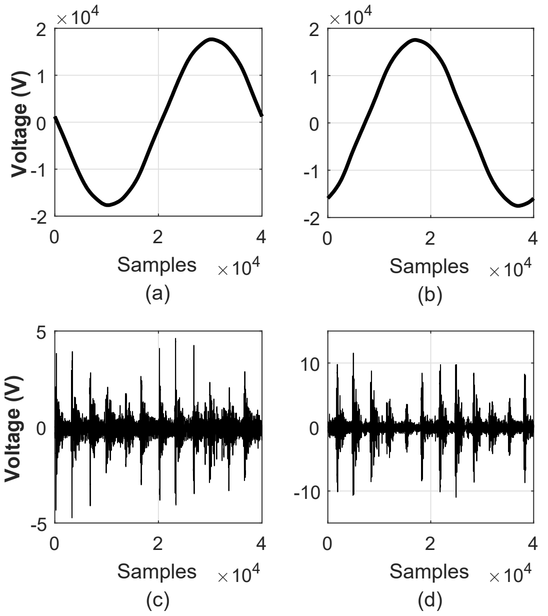


Figure 1: Two example of sampled sweeps. LF recordings from a) Non-fault sweep and b) Faulty sweep. HF recordings from c) Non-fault sweep and d) Faulty sweep.

2.4 Fault signals

The choice made to use sweep sampling mode was probably related to the high sampling rate necessary to characterize the signals in a wide band. The intention to analyze frequencies up to 1 MHz led to a sampling rate of 2 MSa/s, which can generate large amounts of data by the second sampled. Nonetheless, when the trigger signals came to turn on the high sampling recording, both channels (high and low-pass filter) had their signals sampled. This means that despite already being sampled continuously in the 100 kSa/s, the LF signals also had sweeps sampled in the 2 MSa/s sampling rate. That gave the HF sweeps an LF counterpart sampled at the same time, with the same amount of samples. Fig. 1 shows an example of two signals (faulty and not), from sweeps (40 k samples per power cycle) of both channels.

One clear observation from the aforementioned figure is that the fault and non-fault signals are basically indistinguishable. Such illustration is a clear example of the point that such HIFs, with fault currents of single-digit amperes, are much challenging to detect. The main difference between the shown tests is the maximum amplitude of the faulty voltage signal from the HF channel. HF fault signals indeed tend to have higher energy and amplitudes than non-fault ones, but although true, such measure did not show to be consistent enough to be used as a strong predictor. This is mainly because the signals' energy was considerably irregular, not only throughout the test days but also between periods of the day when they were staged.

Although not clearly visible, consistent differences between the two sample stages can be identified. Further investigations showed that the information to distinguish between these two states, however, are probably only to be found in the HF signals.

Beforehand, it is worth to clearly distinguish the difference between LF and HF channels. The characteristic bandwidth of channel connected to the low-pass filter was approximately 5 to 50 kHz, while the high-pass filter connected channel had a 10 kHz to 1 MHz frequency range. The latter is the reason why the amplitude of LF and HF signals on Fig. 1 changes drastically in scale. Therefore, by using the Nyquist-Shannon sampling theorem, it can be concluded that sampling the LF channel at 2 MSa/s was much more than necessary to characterize its related bandwidth. Such conclusion led to the decision to downsample such signals to 100 kSa/s (20 fold) so it could be more numerically efficient.

3 Methodology

This section describes the methodology adopted in the following investigations, which represents a comparative approach between the predictor potential of LF and HF measurements made in the tests signals. This will further presents itself as evidence of the importance of sampling signals at higher rates if accurate detection of these discussed type of faults is to be achieved.

3.1 Measurements

Prior to the brief description of the performed measurements, it is worth stating that the goal of this investigation is not to make inferences regarding the measurements effectiveness at detecting these faults, neither if they are optimal to classify them. This is a comparative analysis that utilizes popular and renowned signal processing techniques to judge the relative information content of LF and HF signals regarding normal and faulty states.

3.1.1 Fourier measurements

Fourier transform is vastly used in signal processing applications, including HIF detection [18, 19]. The technique is used in the present paper to perform Power Spectral Density (PSD) estimation. As shown in Eq (1), the result is basically the squared of the absolute values given by the Fourier transform of the signal. Where, $x[t]$ is the original signal, and $\mathcal{F}\{x[t]\}$ represents its Fourier transform.

$$PSD = \mathcal{F}\{x[t] * x[-t]\} = X[w] \cdot X^{*}[w] = |X[w]|^{2} \tag{1}$$

When applying the Fourier transform to a signal, the outcome will have the same amount of samples as the original signal. Getting the PSD in a useful way means reducing that to half of the samples of the original signal plus one. This results in a feature set with high dimensionality and potential problem in terms of overfitting. Fortunately, as the inferences made here do not rely on the whole set of features, but only the ones that presented the highest separability potential (further explained), having a large number of PSD coefficients will not translate to a severe problem for the comparative analysis. The LF channel, after being down-sampled, resulted in $1k + 1$ features, while the HF channel produced $20k + 1$.

If feature dimension was a problem, however, the strategy would be to derive new features or correlations from the power spectrum components. One of those, the Spectral Flatness (SF) of the signal, was chosen to be compared together with the PSD values. It is also known by tonality coefficient, Wiener entropy or whiteness of a given signal. As described in Eq. (2), where $X[n]$ is the power density spectrum with N number of bins, such measure can characterize the noise-like property of a signal in a zero to one range. It describes how the power spectrum coefficients are distributed in a given bandwidth. For example, perfect sinusoids having no distortion would result in 1, while white noise signals would approach zero.

$$SF = \frac{\sqrt[N]{\prod_{n=1}^{N} X[n]}}{\frac{1}{N}\sum_{n=1}^{N} X[n]} \tag{2}$$

As explicitly depicted, the SF is basically the geometric mean of the PSD values over a giving range (N), divided by its algebraic mean. It is a simple, direct, scale-independent measurement chosen after observing that HIFs tend to create a wide-band noise over the voltages' signals. An interesting note is that it does not necessarily need to be applied to the whole calculated spectrum, but also to arbitrarily different subbands. In the present analysis, the size of these was chosen to be a twentieth of the number of bins of the power spectral density, resulting in 20 features.

3.1.2 Wavelet measurements

Another widely used tool for spectral estimation and signal processing is the Wavelet transform. The Discrete Wavelet Transform (DWT) has a simple, relatively fast, and non-redundant application as shown in Eq. (3).

$$DWT[x, m, n] = \frac{1}{\sqrt{a_o^m}} \sum_{l} x[k] \psi[\frac{n - l a_o^m}{a_o^m}] \tag{3}$$

Where, $x[k]$ is the sampled signal, $\psi[n]$ represents the mother wavelet, a_o^m is the dilation coefficient, and la_o^m is the translation coefficient.

Such approach becomes more numerically efficient when Multi-Resolution Analysis (MRA) is used. In this process, the decomposition is given by the iterative application of a series of low-pass and high-pass filters [20]. The result is time scaled versions of the original signal which have most of its energy in a well-defined bandwidth. Each time a couple of filters is applied, the signal is downsampled in a dyadic manner, resulting in fewer samples, thus giving its numeric advantage. The MRA algorithm is performed as shown in Eq. (4) and (5) by using the $h[t]$ and $g[t]$ as the low-pass and high-pass impulse response functions, respectively.

$$y_d^i[n] = \sum_{k=-\infty}^{\infty} y_a^{i-1}[k] \times h[2n-k] \tag{4}$$

$$y_a^i[n] = \sum_{k=-\infty}^{\infty} y_a^{i-1}[k] \times g[2n-k] \tag{5}$$

The low-pass and high-pass impulse response functions are heavily dependent on the mother wavelet used in the decomposition. A choice for using the same mother wavelet for both signals was made, given the comparative character of the present analysis. The *db*4, a popular mother wavelet from the Daubechies family, with applications in fault detection [21], HIF detection [11, 22], and partial discharge detection [23] was used.

Regarding the measurements, it is important to remember that the outputs of the DWT are still time scaled versions of the original signal that cannot be used directly as features. In this manner, a set of measures was chosen to characterize the outputs of the transformation. They can be listed as: energy percentage (EP), Interquartile range (IQR), L_1, and L_2 norms. The EP is the ratio between the energies of the coefficients and the original signal. The IQR is given by the difference between the upper quartile and lower quartile from signal's distribution, and the L_1 and L_2 norms are the absolute value and Euclidean norm, respectively.

The decision regarding the number of levels used in the decomposition was made by analyzing the lowest (last) frequency band given by the DWT. By applying the MRA algorithm, the upper bound of the bandwidth of a certain level approaches $\frac{F_s}{2^n}$, and the lowest $\frac{F_s}{2^{n+1}}$, where F_s is the sampling frequency and n the decomposition level. This means that a 7-level decomposition would result in an approximation coefficient bandwidth from 0 to 390 Hz, and last detail from 390 to 780 Hz. The authors propose this, by empirical analysis, to be a fair distinction between the low-medium harmonics and other frequency ranges. The same decomposition level was also chosen for the HF signals, which had its last detail bandwidth from 7.8 to 15.6 kHz. This was given as sufficient because this channel had a high-pass filter with

Table 1: Set of measurements' dimensions

		LF	**HF**
Fourier	PSD	1001	20001
	SF	20	20
Wavelet	EP	8	8
	IQR	8	8
	L_1	8	8
	L_2	8	8

10 kHz corner frequency, meaning that investigating lower frequency would not be a useful pursuit.

With all the set of measures described, Table 1 briefly illustrates the dimensionality of the working features. The described scheme resulted in a set of six measurements, two from Fourier domain, and four from the DWT. It might seem a small number of measurements but each one represents a set of multidimensional features. The PSD of the LF signal, for example, results in a feature set of 1001 dimensions. There is an argument for averaging these energy bins as to reduce the number of dimensions, but doing so would result in a loss in frequency resolution. In this configuration, they are separated by values of 50 Hz which are all multiples of the fundamental in the LF channel.

3.1.3 Ranking

In the classic CART (Classification And Regression Tree) algorithm, a measure called Gini Impurity (GI) is used to decide where to split the features dimensions in node creation [24]. In a classification problem, for example, the GI can describe the chance of incorrectly labeling an item in case they were randomly assigned. In CART, one can configure the algorithm to consider every data point of a particular feature as a potential split. When doing so, the GI is used to evaluate each of these splits to find the one that has the highest information gain. The procedure is simple and can be listed as:

1. Select a data point in a particular dimension;
2. Calculate the GI of data pre-split (parent node);
3. Calculate the weighted sum of the GI of both sides of the considered split; and
4. Calculate the GI difference between the pre-split and post-split scenario.

The data point that has the highest GI difference is called the split with the highest information gain. In practice, this will result in a decision boundary that can best distinguish the classification classes when the possibilities of splits are bounded to the data points.

The GI can be calculated by following Eq. (6). Where, J is the number of classes, $i \in \{1, 2, 3, ..., J\}$, p_i is the probability of an observation being corrected labelled in i, and p_k is the probability of mistakenly labelling an observation as in i.

$$GI = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = 1 - \sum_{i=1}^{J} p_i^2 \tag{6}$$

Given that a split is a one-dimensional decision boundary, it becomes possible to find the feature with the higher separability between classes in the measurement set. These can then be ranked and compared between the two types of signals from the LF and HF channels. The ranking of the best splits was performed by their post-split GI (weighted sum), referred to as Impurity Index (I.I.). Therefore, the smaller this index is, more *pure* is the classification zones given by the decision boundary, representing a better discrimination of the data points.

Moreover, in order to add to this comparison, the three best splits in the whole feature set is then used to train a simple decision tree validated by cross-validation. It is noteworthy that the procedure of learning this classifier has only a validation purpose. In this case, it is a way to demonstrate the potential of these features at classifying the signals since the dataset is going to be split into test and training sets, representing the generalization on 'out of sample data'. Considerations regarding overfitting are soon to be discussed.

4 Results

With the aforementioned measurements taken, the single best split (highest information gain) of each set of features is depicted in Table 2 for the LF channel, and in Table 3 for the HF channel. The splits are referred not by their number but by their frequency range or center frequency (PSD) in their respective braces. In the tables, "I.I." stands for Impurity Index (previously discussed) and "Sep." represent the separability potential of the split as a decision boundary. In other words, the separability indicates the percentage of observations that the decision boundary can correctly separate. Such calculation is simply given by the ratio of correct classifications by the total amount of observations. It is similar to the accuracy in case of using such split to classify the whole dataset between the two classes (faulty or not) of sweeps.

Table 2: Split ranking from the LF channel

	Split	I.I.	Sep.
PSD	$PSD\{11.55\ kHz\} > 3.61 \cdot 10^{-6}$	0.47	0.58
SF	$SF\{15 \sim 17.5\ kHz\} > 0.43$	0.45	0.59
EP	$EP\{25 \sim 50\ kHz\} < 3.24 \cdot 10^{-7}$	0.48	0.59
IQR	$IQR\{12.25 \sim 25\ kHz\} > 2.26$	0.43	0.64
L_1	$L_1\{12.5 \sim 25\ kHz\} > 985.85$	0.44	0.63
L_2	$L_2\{25 \sim 50\ kHz\} > 22.98$	0.45	0.62

Table 3: Split ranking from the HF channel

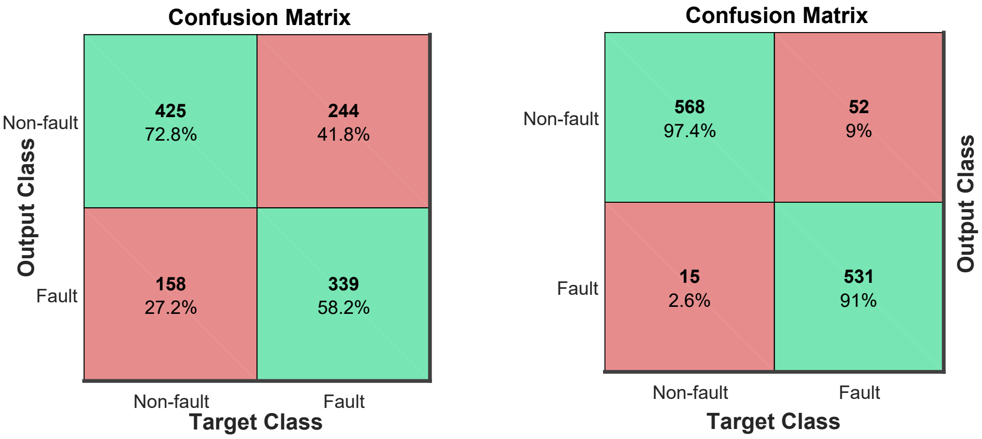
	Split	I.I.	Sep.
PSD	$PSD\{34.5\ kHz\} > 5.12 \cdot 10^{-6}$	0.4	0.68
SF	$SF\{1 \sim 50\ kHz\} > 0.02$	0.42	0.66
EP	$EP\{31.25 \sim 62.5\ kHz\} > 1.21$	0.39	0.69
IQR	$\mathbf{IQR\{62.5 \sim 125\ kHz\} > 0.05}$	**0.15**	**0.91**
L_1	$L_1\{62.5 \sim 125\ kHz\} > 250.02$	0.22	0.87
L_2	$L_2\{15.62 \sim 31.25\ kHz\} > 16.67$	0.32	0.77

It may be worth noting that the tables do not discriminate between faulty or non-faulty observations. The I.I. and Sep. need to be calculated considering the whole data set to make sense.

Such results can demonstrate the difference between the information content regarding vegetation HIF discrimination in both signals in a comparative manner. The impurity index shows that the HF measurements overperform the LF features, although close in some splits, at every comparison. In the same manner, the separability showed that such measurements in the LF channel, when used as predictors to classify such faults, is not much reliable than a coin toss at labelling the observations.

It is important to notice that better feature extraction methods can, and probably would, result in higher discrimination values for both channels. Also, that only two features extracted at the HF channel indicated reasonable decision boundaries for fault occurrences separability, namely one IQR and one L_1 measure. Nevertheless, it does not necessarily take the value of the comparative analysis away. Such improvement, gained by better feature extraction and hypothesis set for classifying the faults, was attested in previous work by the authors in [6]. The paper demonstrates how a combination of Wavelet measurements used in an ensemble of decision trees can result in accuracy values greater than 98%, with 99% of overall security.

Regarding the best predictor from the performed experiments, the Inter-



(a) LF tree's confusion matrix

(b) HF tree's confusion matrix

Figure 2: Confusion matrixes from the best three splits in both LF and HF channels.

quantile Range from the HF channel showed the lowest impurity and higher significance. With an impurity index of 0.15 and separability of 0.91, it represented a promising feature from the studied type of faults. Another worth mentioning observation is the superiority of DWT features over the Fourier measurements for both channels. This has been a consensus in the literature [22, 25, 26], which confirms the Wavelet transform ability to better represent fast transients in the fault signals.

Having separability of 0.91 means that if used as a stand-alone predictor, such feature would correctly separate 91% of the dataset samples. However, separating a whole dataset is not evidence for generalization, i.e., correctly predicting new data (out of sample observations). To address such issue, inspired by previous results shown in [6], a further comparative experiment was made. The three best splits over the whole set of features of each channel were select to fit a simple decision tree, validated in 2-fold cross-validation. This means dividing the dataset in two, using half 1 to learn the tree, and half 2 to test it. Also, doing it the other way around (fitting with 2 and testing with 1), and reporting the average out of sample error. In this experiment, to illustrate the effect of choosing a different current threshold, a 0.5 A value was used rather than the 0.1 A previously chosen.

The confusion matrix, a well-known performance illustrator of machine learning algorithms resulted from classifiers learned from both channels, is given in Fig. 2. Note the reducing of observations numbers used in each class from 768 to 583. This is due to the fact that a good portion of the tests did not reached 0.5 A.

In the confusion matrix representation, the green squares attest for accurate classification, while the red for misclassified observations. The lines represent the output classification, while the columns attest to the target (actual) class. For example, an element in line 2 and column 1 represents the number of observation classified as faults (output), that was actually a Non-fault observation (misclassification).

When considered by overall accuracy, the simple tree fitted with three HF features correctly classified 94.2% of the observations, while the tree with LF features only labeled 65.5% of observations correctly.

To illustrate the promising results of such approach, Fig. 3 depicts the voltage, current, and classifier output of test #504. This experiment is part of an exceptional group of tests, composed by a few recordings that had 'pre-fault' recordings. The voltage starts moments before the 5^{th} second but current conduction starts close to the 8^{th} second. Current reaches the 0.1 A threshold value between the 10^{th} and 11^{th} second, and never exceeds 2 A in amplitude. The verticals lines in the output bounds the period where sweeps are considered to be 'in fault'. In case this test was used for training, only the 11^{th} sweep would be selected (first after threshold reached). However, correct classification starts in a prior sweep, recorded at the 10^{th} second. When the fault was interrupted after the $16th$ second, the classifier was able to reset itself.

To avoid any confusion, it may be worth mentioning that although the signals illustrated in Fig. 3 came from the LF channel, the classification was made based on the HF features. HF signals were not illustrated because they are small sweeps per second that do not properly represent the test when concatenated in a timeline. LF signals, sampled in the continuous method, clearly illustrate the evolution of the fault current and its inception.

5 Discussions

The fact that high accuracy was achieved by such simple decision boundaries is noteworthy. As commonly known in the machine learning field, the chances of the results not generalizing in out of samples observations increase with the complexity of the utilized hypothesis set (Vapnik–Chervonenkis dimension). This is avoided by the use of small number of features and cross-validation. The results not only showed to be promising to practical applications but also that there is much to be investigated in terms of phenomena understanding when it comes to vegetation HIFs. Equally important, the fact the promising results were achieved by using 'weak' features (predictors), far away from being the optimal ones, only compounds on this observation.

One may be skeptical regarding the need for this level of current sensitivity or the fact that such sampling is ever going to be practical. Regarding its need, however, one can find strong supporting evidence in the current literature. A recent paper [10] that deliveries an analysis of the fires dataset

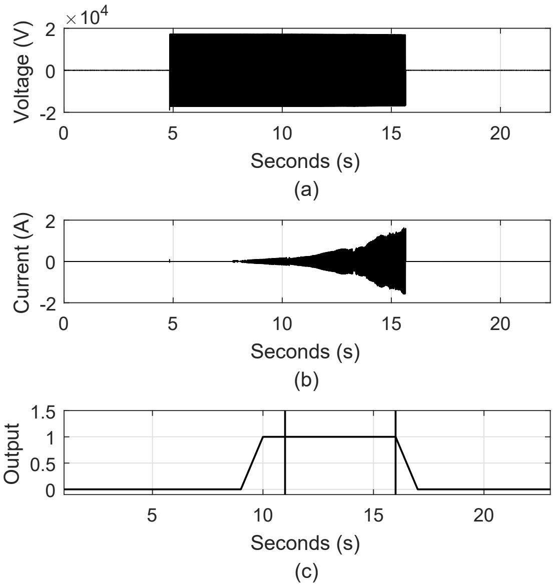


Figure 3: Example of classification potential. a) Recording of the voltage test by the LF channel, b) Current signal, and c) Classifier output.

in the state of Victoria and their causes can be cited as an example. The presented results point out that although the number of fire events caused by electric distribution assets is over-represented, they are responsible for creating the greatest amount of damage. The findings of [9] also support such need by stating that vegetation HIFs can ignite fires even when they result in single-digit current amplitudes. And finally, the lack of evidence in recent literature for limited current vegetation HIFs detection also strengthens such need if fire risk is to be reduced. Regarding practicability, there are relevant arguments that point this to be a technology worth investing. The more important, however, may be that such high sampling technology will probably not be related to existing hardware, such as substation voltage transformers, due to economic costs. Rather, the solution will come from distance measures made by novel sensor technologies. In fact, recent patents publications [27, 28] show the increasing interest in the application of such solutions and its possible more affordable commercialization. Document [28] disclosures a solid-state electric-field sensor that can sample signals for a target located at distance (powerlines) by the change in the electric field in a voltage-controlled capacitor. In the same manner, [27] discloses an early fault detection system that samples signals with antenna sensors, also without the need for physical connection (at distance). The increasing interest in these sensor technologies is the reason why the voltage signals were previously claimed as the domain of interest to tackle the vegetation

HIFs, rather than current signals. Yet, further reasonable points could also be cited to highlight the interest for the similar technologies. For instance, the fact that it is highly probable that more distributed and sophisticated sampling methods are going to be needed once distribution systems evolve towards a smart grid scenario. The possibility to aid other problems and disturbances diagnostics that could benefit from such measurements. For example, the standing problem of accurate fault location and power quality estimation, which increases in complexity with the growing penetration of distributed renewable generation.

6 Conclusions

This paper presents a case for the need of a more comprehensive data sampling to address the vegetation HIF detection problem. The arguments are mainly inspired by the lack of consensus regarding a practical solution in the related literature, despite its lasting history, and the complexity of the problem itself. The case was presented by a set of experiments performed in a dataset of real staged faults regarding the signals' information content to reliably discriminate fault occurrences. Results indicated that the information to differentiate between a normal and faulty state is probably only to be reliably found in the signals' high-frequency components. This was demonstrated by the calculation of the Impurity Index and Separability for both sampling methods presented. The outcomes add to aforementioned cited arguments that such faults should receive special attention due to their specific fault signatures and fault currents. The fact that Fourier and Wavelet measurements were used in the analysis also endorsed, as additional value, the consensus of Wavelet transform being advantageous in HIF disturbances representations. And finally, reasons to regard the importance and feasibility of using such technology were discussed, corroborating its need for the mitigation of fire risk created by powerlines.

Acknowledgement

The authors would like to thank the Powerline Bushfire Safety Program for funding such massive staged faults program and making resulted data available for this research. This work was supported in part by the Department of Economic Development, Jobs, Transport and Resources from Victoria, Australia.

Declaration of interest

The authors do not have any conflict of interest to declare.

References

[1] B. M. Aucoin and B. D. Russell, "Distribution high impedance fault detection utilizing high frequency current components," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, no. 6, pp. 1596–1606, 1982.

[2] C. J. Kim, B. D. Russell, and K. Watson, "A parameter-based process for selecting high impedance fault detection techniques using decision making under incomplete knowledge," *IEEE Transactions on Power Delivery*, vol. 5, no. 3, pp. 1314–1320, 1990.

[3] K. Pandakov, H. K. Høidalen, and J. I. Marvik, "Misoperation analysis of steady-state and transient methods on earth fault locating in compensated distribution networks," *Sustainable Energy, Grids and Networks*, 2017.

[4] J. Vico, M. Adamiak, C. Wester, and A. Kulshrestha, "High impedance fault detection on rural electric distribution systems," in *2010 IEEE Rural Electric Power Conference (REPC)*, 2013, Conference Proceedings, pp. B3–B3–8.

[5] D. Hou, "High-impedance fault detection—field tests and dependability analysis," in *Proceedings of the 36th Annual Western Protective Relay Conference, Spokane, WA*, 2009, Conference Proceedings.

[6] D. P. S. Gomes, C. Ozansoy, and A. Ulhaq, "High-sensitivity vegetation high impedance fault detection based on signal's high-frequency contents," *IEEE Transactions on Power Delivery*, vol. 33, pp. 1398–1407, 2018.

[7] N. Bahador, F. Namdari, and H. R. Matinfar, "Modelling and detection of live tree-related high impedance fault in distribution systems," *IET Generation, Transmission & Distribution*, vol. 12, no. 3, pp. 756–766, 2018.

[8] ——, "Tree-related high impedance fault location using phase shift measurement of high frequency magnetic field," *International Journal of Electrical Power & Energy Systems*, vol. 100, pp. 531–539, 2018.

[9] T. Marxsen, "Vegetation Conduction Ignition Test Report - Final," Marxsen Consulting Pty Ltd., Department of Economic Development Jobs Transport and Resources, 2015.

[10] C. Miller, M. Plucinski, A. Sullivan, A. Stephenson, C. Huston, K. Charman, M. Prakash, and S. Dunstall, "Electrically caused wildfires in Victoria, Australia are over-represented when fire danger is elevated," *Landscape and Urban Planning*, vol. 167, pp. 267–274, 2017.

[11] W. C. Santos, F. V. Lopes, N. S. D. Brito, and B. A. Souza, "High-impedance fault identification on distribution networks," *IEEE Transactions on Power Delivery*, vol. 32, no. 1, pp. 23–32, 2017.

[12] Bureau of Meteorology. Australian Government, "Understanding fire weather," 2017. [Online]. Available: http://media.bom.gov.au/social/blog/1538/understanding-fire-weather

[13] Victoria State Government, "Powerline bushfire safety program," 2018. [Online]. Available: http://earthresources.vic.gov.au/energy/safety-and-emergencies/powerline-bushfire-safety-program

[14] A. Ghaderi, H. L. Ginn Iii, and H. A. Mohammadpour, "High impedance fault detection: A review," *Electric Power Systems Research*, vol. 143, pp. 376–388, 2017.

[15] 2009 Victorian Bushfires Royal Commission, "Final report," Victoria State Government, Report, 2010. [Online]. Available: http://www.royalcommission.vic.gov.au/finaldocuments/summary/PF/VBRC_Summary_PF.pdf

[16] M. Sarlak and S. M. Shahrtash, "High-impedance faulted branch identification using magnetic-field signature analysis," *IEEE Transactions on Power Delivery*, vol. 28, no. 1, pp. 67–74, 2013.

[17] J. C. Chen, B. T. Phung, D. M. Zhang, T. Blackburn, and E. Ambikairajah, "Study on high impedance fault arcing current characteristics," in *2013 Australasian Universities Power Engineering Conference (AUPEC)*, 2013, Conference Proceedings, pp. 1–6.

[18] A. Soheili, J. Sadeh, and R. Bakhshi, "Modified FFT based high impedance fault detection technique considering distribution non-linear loads: Simulation and experimental data analysis," *International Journal of Electrical Power & Energy Systems*, vol. 94, pp. 124–140, 2018.

[19] V. Torres, J. L. Guardado, H. F. Ruiz, and S. Maximov, "Modeling and detection of high impedance faults," *International Journal of Electrical Power & Energy Systems*, vol. 61, pp. 163–172, 2014.

[20] S. G. Mallat, "A theory for multiresolution signal decomposition: the Wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[21] K. M. Silva, B. A. Souza, and N. S. D. Brito, "Fault detection and classification in transmission lines based on Wavelet transform and ANN," *IEEE Transactions on Power Delivery*, vol. 21, no. 4, pp. 2058–2063, 2006.

[22] J. Chen, E. Ambikairajah, D. Zhang, T. Phung, and T. Blackburn, "Detection of high impedance faults using current transformers for sensing and identification based on features extracted using Wavelet transform," *IET Generation, Transmission & Distribution*, vol. 10, no. 12, pp. 2990–2998, 2016.

[23] X. Ma, C. Zhou, and I. J. Kemp, "Interpretation of Wavelet analysis and its application in partial discharge detection," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 9, no. 3, pp. 446–457, 2002.

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees. wadsworth & brooks," *Monterey, CA*, 1984.

[25] S. H. Mortazavi, Z. Moravej, and S. M. Shahrtash, "A hybrid method for arcing faults detection in large distribution networks," *International Journal of Electrical Power & Energy Systems*, vol. 94, pp. 141–150, 2018.

[26] K. Chul-Hwan, K. Hyun, K. Young-Hun, B. Sung-Hyun, R. K. Aggarwal, and A. T. Johns, "A novel fault-detection technique of high-impedance arcing faults in transmission lines using the Wavelet transform," *IEEE Transactions on Power Delivery*, vol. 17, no. 4, pp. 921–929, 2002.

[27] K. L. Wong and A. Bojovschi, "Fault detection system," U.S. Patent US9 606 164B2, mar 28, 2017.

[28] M. A. Noras, "Solid-state electric-field sensor," U.S. Patent US9 846 024B1, dez 19, 2017.