# Early Detection of Vegetation Ignition Due to Powerline Faults

Sevvandi Kandanaarachchi, Nandini Anantharama, and Mario A. Muñoz

*Abstract*—**High impedance faults through contact with vegetation are one of the main causes of electrically caused wildfires. While detecting these faults is challenging on its own, it is important to do so in the context of the risk of vegetation ignition, as disconnecting the power infrastructure can have unwarranted, damaging consequences during an emergency. Hence, we propose a methodology for prevention of wildfires, through the accurate prediction and early detection of the ignition risk resulting from high impedance faults. Our methodology uses a set of features derived from time- and frequency-domain analyses. To test our methodology, we use a large, publicly available experimental dataset. Our results demonstrate that the methodology allows the detection of ignition risk well before its onset with high accuracy.**

*Index Terms*—**Anomalous time series detection, fire-risk modelling, high-impedance faults, vegetation faults, wildfire ignition.**

## I. Introduction

FAILURES in electrical infrastructure have been at the centre of severe wildfires across the world, particularly in the United States, Spain and Australia [1]. Powerline faults in specific can start fires through three well-known ignition mechanisms: incandescent metal particles emitted when high voltage conductors clash; high voltage arcs that occur near vegetation; and high voltage current that passes through vegetation [2]. Although powerline faults are not one of the main reasons why wildfires start, e.g., in Victoria, Australia, 1.5–3.0% of all fires between 2007 and 2014 were due to electrical sources [3]; these tend to occur during elevated fire danger conditions, resulting in quick spread and more severe consequences [1]. In fact, the average size of a large electrical-caused wildfire is an order of magnitude higher than large fires due to other causes [1].

High voltage current that passes through vegetation, unlike the other two ignition mechanisms, is not well understood [2]. This is because it is a High-Impedance Fault (HIF): an electric fault whose current amplitude does not exceed the threshold of protection devices. As such, they do not represent stress risk to the equipment and can easily be confused by increased customer load [4]. Ghaderi *et al.* [5] and Jazebi *et al.* [6], [7] provide comprehensive reviews of HIF detection approaches. Typically, HIF detection algorithms rely on arcing fault signatures. For example, Zhang and Jing [8] proposed a generic model for detecting both high currents and high impedance arc faults. Their approach evaluates the fault signature for best fit with an arc fault model and a non-arcing disturbance model, and provides a binary outcome for identifying arc faults. Mukherjee *et al.* [9] proposed an electromagnetic radiation (EMR) centred approach for the detection of arcing faults in low voltage distribution systems. This approach uses the Log-Spectral distance metric of EMR sensor data to detect arcing faults. While the algorithm is able to detect arcing faults as well as distinguish between arcing faults and arc-mimicking faults, it is sensitive to the distance of the sensor from the arc current path, and does not yet support identification of the source.

Vegetation HIFs are caused by powerlines breaking and falling onto vegetation at ground level, vegetation brought by heavy winds bridging two phase conductors, or tall trees reaching powerlines [5]. Unlike arcing faults, vegetation HIFs differ in terms of lower frequency components and growth rate of current [7]. Moreover, vegetation HIFs are influenced by factors such as grounding type, voltage level, fault impedance value, signal sampling parameters, and the characteristics of the contact surface, among others [4]. For example, a study by Guggenmoos [10] identifies off-row trees as a major source of tree related outages, and provides a model to quantify tree failure risk with a trade-off between line risk and clear width. The trade-off is provided as a cost-benefit representation, thus enabling an informed comparison of benefits of clearance width with other alternatives. Due to the diversity of factors, detecting Vegetation HIFs is a challenge. Previous analyses that have been effective in identifying these type of faults use electrical conductivity of the vegetation's sap and risk level of the vegetation area as influencing factors [11], [12].

Gomes *et al.* [4], [13], [14] have focused on the characterization of vegetation HIFs, particularly by examining the fault signatures' high frequency content. In their work, they employed the real-world Powerline Bushfire Safety Program (PBSP) dataset comprising a large number of experiments,

Manuscript received January 30, 2020; revised April 16, 2020 and June 4, 2020; accepted June 22, 2020. Date of publication July 2, 2020; date of current version May 21, 2021. This work was supported in part by the Australian Research Council under Grant LP160101885 and in part by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) through their research support scheme. Paper no. TPWRD-00123-2020. *(Corresponding author: Sevvandi Kandanaarachchi.)*

Sevvandi Kandanaarachchi is with the School of Science, Mathematical Sciences, RMIT University, Melbourne, VIC 3000, Australia, and also a Research Affiliate with the Department of Econometrics and Business Statistics, Monash University, Clayton, VIC 3800, Australia (e-mail: sevvandi.kandanaarachchi@rmit.edu.au).

Nandini Anantharama is with the Faculty of Information Technology, Monash Unviersity, Clayton, VIC 3800, Australia (e-mail: nandini.anantharama@monash.edu).

Mario A. Muñoz is with the School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia (e-mail: munoz.m@unimelb.edu.au).

Digital Object Identifier 10.1109/TPWRD.2020.3006553

sampled in a functioning network in the presence of noise [2]. From this dataset and each fault they collected one sample of 20 ms from a high frequency voltage signal, i.e., in the 0 kHz to 50 kHz frequency range and the 10 kHz to 1 MHz bandwidth. Using Fourier [4] and Wavelet [13] transforms, and Shift-Invariant Sparse Coding [14] to characterise the signal patterns, Gomes *et al.* used decision tree classifiers to separate faulty from non-faulty signals, with the aim of disconnecting power immediately after detecting a fault. Their results achieved a high detection performance, with a true positive rate of 97.4%, confirming that the signatures' high frequency content help discriminate vegetation HIFs.

However, while characterizing and detecting these HIFs is challenging on its own, it is important to do so in the context of the risk of vegetation ignition. Without appropriate consideration given to such risk, disconnecting power when faults are detected is sub-optimal. Moreover, it can have unwarranted, damaging consequences. For example, during emergencies, loss of power will disrupt among other essential services: communication systems, affecting the dissemination of important information; traffic signaling, increasing the risk of road accidents; water distribution, hampering fire fighting activities; and air cooling systems, increasing the likelihood of heat related sicknesses [15].

Therefore, in this work we propose a methodology for prevention of wildfires, through the accurate prediction and early detection of the ignition risk resulting from HIFs. To test our methodology, we make use of the PBSP dataset [2], previously employed by Gomes *et al.* in their work [4], [13], [14]. This dataset is encoded in the proprietary `.pnrf` file format. Therefore, we also provide the software package `pnrfr` for the R programming language [16], which besides importing seamlessly the PBSP data and providing other basic operations, facilitates the access to the vast array of statistical and machine learning methods available in R to other researchers using this dataset.

The remainder of the paper is organised as follows. In Section II, we describe the data employed in this research. Then, in Section III, we describe our methodology to pre-process, characterise and classify the fault signatures depending on their risk of ignition. In Section IV, we present the results of our experiments. We finalise this paper with our conclusions in Section V.

## II. Background

### A. Black Saturday Bushfires

The "Black Saturday bushfires" were a series of catastrophic wildfires that occurred in the state of Victoria, Australia, in February 2009, when temperatures often exceeded 45 °C. Collectively, they killed 173 people, injured over 4000 more, burnt over 270000 ha, and caused over $4.4 billion in economic damages to the state, including the destruction of 1832 homes [2], [17]. Five of the eleven major wildfires were caused by electricity assets, accounting for over 70% of the fatalities and 60% of the burnt area. Moreover, two of them, known as the Beechworth-Mudgegonga and Coleraine bushfires, were directly caused by vegetation contact with powerlines accounting for two fatalities and 13% of the burnt area [17]. The state of Victoria has a record of electrically caused, catastrophic wildfires. For example, in the months of February of 1977 and 1983, both known as devastating seasons, over half of the major wildfires were caused by electrical infrastructure [17]. Therefore, in the aftermath of the "Black Saturday bushfires," an inquiry known as the 2009 Victorian Bushfires Royal Commission was carried out. Some of its recommendations focused on improved management of electrical infrastructure, particularly on days of peak fire danger, and requirements for distribution business and municipal councils to identify and reduce the risk posed by "hazard trees," i.e., trees that are outside the clearance zone but that could come into contact with an electric powerline having regard to foreseeable local conditions [17]. For this purpose, the electricity state regulator oversees pruning programs that that operate on a two to three year cycles, and fosters community engagement for the reporting of tree clearance issues [18]. However, due to the vast amounts of uninhabited forestland, pruning may not be frequent enough to detect all "hazard trees," making essential the development of automated monitoring equipment and tools that facilitate the rapid identification of problem areas, and disconnection of a powerline if required.

Besides the response by the state regulators, the findings of the Commission have also resulted in research being carried out from different disciplines. For example, Di Giulio [19] discusses the policy implications concerning the replacement of Single Wire Earth Return powerlines. Oloruntoba [20] explores the challenges wildfires poses to disaster planning, and effective strategies for emergency response planning. Eburn *et al.* [21] discuss new approaches for post-event reviews as alternatives to government enquiries such as a Royal Commission. Williamson [22] discusses the role of solar photo-voltaics and energy storage systems in mitigating bushfires. Roozbahani *et al.* [23] discuss a mathematically based optimisation approach for asset replacement as an alternative to the current practice based on expert interpretation of risk maps. Whittaker *et al.* [24] explore the factors affecting the severity of bushfires and discuss ways to adapt to a changing climate. Miller *et al.* [1] explore the extent of damage caused by electrically caused bushfires. Broome *et al.* [15] investigate the health risks arising from cutting off power during periods of high bushfire threats, and argue that cutting off power leads to more deaths and higher costs to communities. Gomes *et al.* [4], [13], [14] focus on the detection of high impedance current faults, particularly by characterising the fault signatures' high frequency content. However, none of this research proposes methods for safely disconnecting power after a fault considering the risk of vegetation ignition.

The Victorian Government also responded to the findings by establishing the Powerline Bushfire Safety Program (PBSP), which focused on research to improve the knowledge on how fires can start from powerline faults and how they might be prevented [2]. The two main goals from the PBSP were to: (a) identify the worst species for fire starts from powerline faults and understand their ignition processes; and (b) to compile a reference data base of electrical signals caused by vegetation faults to support development of better fault detection technology. To fulfill its second aim, the PBSP commissioned a series of tests which resulted in the vegetation fault database described in the following section. Like other anomaly detection problems such as intrusion detection, cyber crime, and terrorist attacks, where uncommon but extremely damaging events must be avoided, the PBSP database provides a benchmark for testing essential technology, for which we cannot wait for real-world data. This is despite testing being carried out under a controlled environment, which may not perfectly mirror real-world conditions.

### B. The PBSP Vegetation Fault Database

During the summer season of 2014-15, a test rig was constructed inside an insulated shipping container, where extreme fire weather conditions such as dry wind were simulated [2]. The rig was connected to a real network via current-limiting resistors. Stewart [25] discusses the design and operation of the test facility that was used to conduct over 1000 tests aimed at testing various HIFs and the associated risk of fire in a safe environment.

Voltage and current signals were collected using low-noise and wide-bandwidth measuring systems at two different sample rates, i.e., *low* frequency data was continuously sampled at 100 kHz, while *high* frequency data was sampled at 20 ms/s bursts at 2 MHz. Vegetation samples of various sizes from 18 species were prepared by drying them at 45 °C for up to 24 h. The selected species are mostly native and unique to Australia, with some exceptions such as *Acacia melanoxylon* and *Fraxinus angustifolia*, which are also present in Africa, Asia and Europe; *Cotoneaster glaucophyllus* is native to China and the Himalayas region; and *Schinus molle* is native to South America. The tests yielded a database of fault signatures composed of 1038 fault tests and 112 calibration/noise tests. Three fault geometries were used:

1) **Phase-to-earth** (P2E – 389 valid tests), where a branch was laid across two high voltage conductors, one of them earthed and the other one with full nominal phase voltage applied, i.e., 12.7 kV between them.
2) **Phase-to-phase** (P2P – 389 valid tests), where a branch was laid across two high voltage conductors connected to separate phases of the incoming high voltage supply, i.e., 22 kV between them. These tests resulting in faults with a growth rate four times faster than P2E, with an initial current also 70% to 80% higher.
2) **Wire-into-vegetation** (W2V – 260 valid tests), where a high voltage conductor energised at 12.7 kV was placed into earthed vegetation, either grass or bush.

Fig. 1 illustrates a 10 ms rolling window RMS current signal during a vegetation HIF. Where the vegetation goes through irreversible physical changes, the fault current tends to take up where it left off [2]. Four phases of the failure are presented [2]:

1) **Development of conductor-vegetation contact** which lasts between the application of high voltage to the vegetation sample until the first maximum peak current value. During this phase, small embers develop. Moreover, conductive plasma and flame grow in the contact points, which increase the overall contact area. This phase usually takes between 10 s to 15 s. However, it can take more than a minute in extreme cases. During this phase, current may exceed the protection levels – often set at either 0.5 A, 1.0 A, 2.0 A, 4.0 A – disconnecting the powerline.
2) **Expulsion of moisture** which lasts from the first maximum peak current value until the next minimum peak. During this phase, the sample released steam and water, often accompanied by loud noises. Tests that went on into this phase often continued into the next.
3) **Progressive charring extension** which last until the ultimate appearance of runaway current. During this phase, flame slowly spreads along the sample, producing intermittent arcs that would briefly short-circuit the system, creating large current fluctuations.
4) **Flashover** which occurred when the maximum current value appeared – 45.0 A in phase-to-phase tests and 65.0 A in phase-to-earth tests – due to the flame extending from conductor to conductor creating a unbroken short-circuiting path.

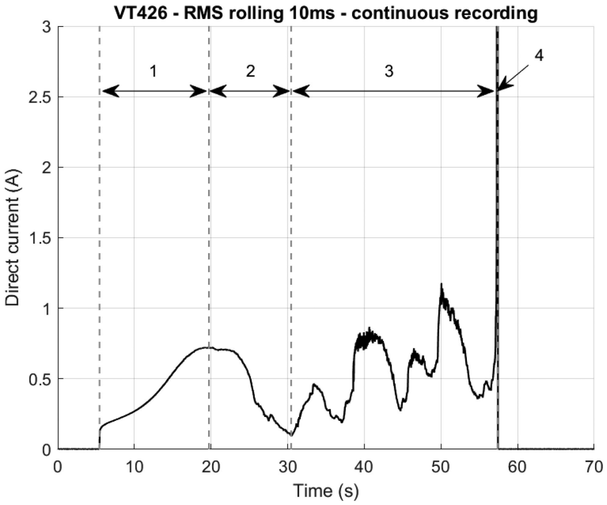

Fig. 1. LF current signal (RMS) of test VT426, which illustrates the four phases of the vegetation high impedance fault: (1) Development of conductor-vegetation contact; (2) Expulsion of moisture; (3) Progressive charring extension; and (4) Flashover. The probability of ignition should be established during phase 1.

These tests also concluded that samples with moisture content below 10% to 15% do not conduct enough current to cause thermal runaway; larger samples drew high levels of initial current, which quickly increased to reach the pre-set limits; and most conductive layers of the sample were those immediately under the bark [2].

## III. Methodology

Our ignition risk detection methodology can be divided in four steps, i.e., importation, pre-processing, feature calculation and classification. The details of each step are described in the following sections.

### A. Data Importation

At the core of the PBSP vegetation fault database is the collection of signatures recorded in the proprietary *Perception Native Recording Format* `.pnrf` by HBM. While broadly used

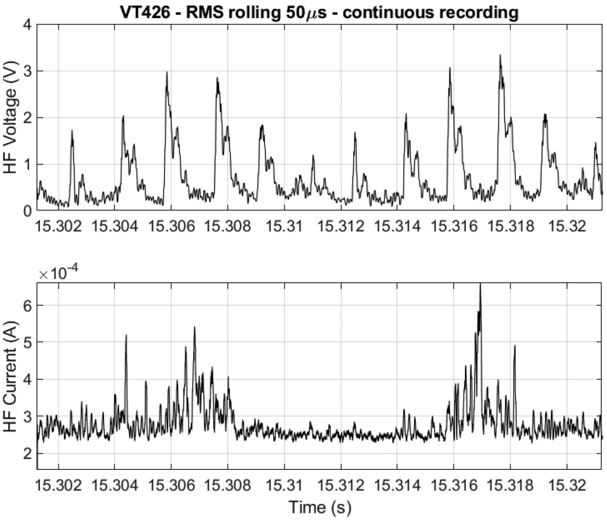

Fig. 2. RMS signals from a 50 $\mu$s burst of high-frequency data during test VT426.

in data acquisition, as it efficiently stores multi-channel data records, this format is unfamiliar to the broader Machine Learning community. As such, tools to read this format in R or Python, the two most commonly used languages by this community, are not available. Therefore, the researchers working with the PBSP database would not have access to a broad array of state-of-the-art methods for classifying complex data such as this.

Hence, we provide the `pnrfr` package [16] as a R wrapper around the PNRF Toolkit Reader software by HBM. While limited to the Windows Operating System at the moment, this package provides an R interface to read PNRF channel data. Moreover, it provides access to meta information such as name, type, unit of recording and the sampling interval for each channel in the file. Finally, it provides some additional functionalities such as down sampling the data and computing RMS value for a given window period.

### B. Data Pre-Processing

As described in Section II, each one of the valid tests has four channels of signal data, i.e., voltage and current both sampled at low (100 kHz) and high (2 MHz) frequencies. Low frequency (LF) signals were continuously recorded, while high frequency (HF) signals were recorded at 20 ms/s bursts, resulting in an equal number of samples per second on all channels. We computed a RMS signal by using a rolling window of 10 ms for the LF channels, and a 50 $\mu$s one for the HF channels. Fig. 2 illustrates the RMS signals for one 20 ms burst of HF signals.

The signals were recorded in a variety of conditions. For example, recording started with and without voltage being applied. As such, we define starting point of the test as the first sample from the HF signals, for which the RMS LF Voltage is higher than 10 kV and the RMS LF current is higher than 100 mA. The end point of the test was defined as the first peak of the RMS LF current using a rolling window of 500 ms, which corresponds to the end Phase 1. According to Marxen [2], the onset of ignition occurs at some instant during Phase 2 onward, whose exact value is unknown. Given that we are interested in characterising the risk of ignition, we examine the signals until the risk is at its highest. To identify the first peak, we employed the `findpeaks` algorithm from MATLAB. The four raw signals and four RMS signals were then divided into non-overlapping windows of $100 \times 10^3$ samples.

### C. Features

Marxen [2] considers a combination of non-linear time- and frequency-domain techniques as the most promising approach to detect the disturbance created by a vegetation HIFs. Previous studies [13], [14] have focused on frequency-domain techniques. Hence, we use a simple, low-cost approach based on features to characterize the signals in both domains. This approach is popular in time-series related tasks, such as forecasting [26], classification [27], clustering [28], similarity queries [29] and anomaly detection [30]. In our approach, the time-domain analysis is performed in both the raw and RMS signals. However, the frequency-domain analysis is only performed on raw data, as the RMS signal has filtered out most high-frequency components. The details are presented in the following sections.

*1) Time-Domain Features:* We computed a total of 112 time-domain features for each window. These features can be categorized into four types: STL, statistical, auto-correlation and current acceleration. STL stands for Seasonal and Trend decomposition using LOESS (LOcally Estimated Scatter-plot Smoothing) [31]. From each one of the eight signals, we compute 13 features, i.e., 4 STL, 3 statistical and 6 auto-correlation features, giving a total of 104. Eight additional current acceleration features are taken from the RMS LF current only. The first three types were calculated using the R package `tsfeatures` [32], while the last type was developed from consultation with the PBSP staff. In the following sections we describe the details of each feature.

*a) STL decomposition based features:* The STL decomposition takes the original signal $x_t$, and identifies three components that represent certain behaviours. These are the *trend* component, $f_t$, which represents the long-term progression of the signal and any repeated but non-periodic fluctuation; the *seasonal* component, $s_t$, which represents a periodic fluctuation; and the *remainder* component, $e_t$, which is expected to be the normally distributed random noise that remains after the other two components have been removed. The original signal can be reconstructed by adding these three components, i.e., $x_t = f_t + s_t + e_t$. Fig. 2 illustrates the STL decomposition of the HF Voltage signal first illustrated in Fig. 2. Once decomposed, the 'trend' feature is calculated as follows:

$$\text{trend} = \max\left(0,\, \min\left(1,\, 1 - \frac{\operatorname{var}(e_t)}{\operatorname{var}(f_t + e_t)}\right)\right), \qquad (1)$$

where var denotes the variance. As such, 'trend' lies within the $[0,\ 1]$ range, with '0' indicating a weak trend and '1' a strong one.

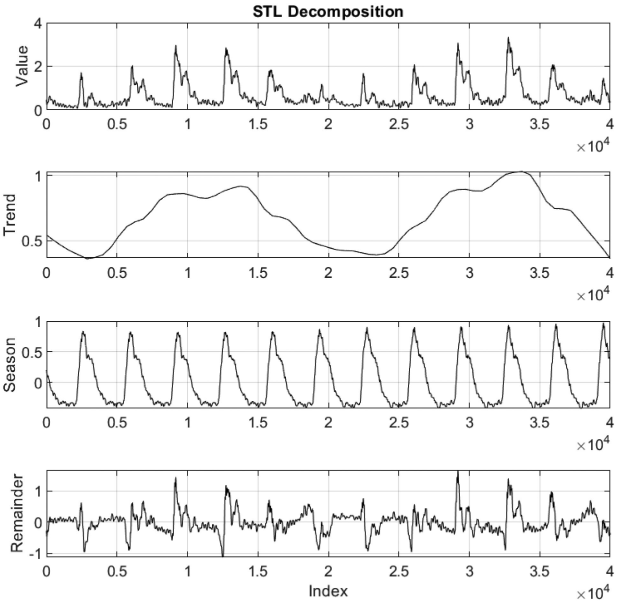

Fig. 3. STL decomposition of the RMS HF Voltage component from of test VT426, initially illustrated in Fig. 2.

The ‘spike’ feature is a measure of the abrupt deviations from the normal values. It is calculated using the Leave-one-out (LOO) method applied to the remainder component. In other words, from a signal $\{e_1, \ldots, e_t, \ldots, e_N\}$, we create $N$ new signals each one missing an entry. For example, we denote LOO$(e_t)$ as the signal missing $e_t$. Let $v_{e_t}$ be the variance of LOO$(e_t)$. Then ‘spike’ is defined as:

$$\text{spike} = \text{var}\left(\{v_{e_t}\}_{t=1}^{N}\right) . \tag{2}$$

High ‘spike’ values suggest the occurrence of obvious spikes in the time series, while low values indicate non-spiked data.

Two more features, ‘e_acf1’ and ‘e_acf10,’ are derived from an auto-correlation analysis of the remainder component, which aims to capture any additional structure as observed in Fig. 3. The former correspond to the first auto-correlation coefficient, i.e., between $e_t$ and $e_{t-1}$, while the latter is the sum of the first ten squared auto-correlation coefficients. Low values of both ‘e_acf1’ and ‘e_acf10’ suggest that $e_t$ is similar to white noise, while high values indicate the existence of structures not captured already by the STL decomposition.

*b) Features derived from statistical analysis:* The second type of features are derived from statistical analysis of $x_t$. Three features are obtained through this approach. These are ‘curvature,’ ‘linearity’ and ‘entropy’. The first two are second and first order coefficients resulting from fitting a quadratic polynomial model to the signal, i.e., $x_t = \alpha_2 t^2 + \alpha_1 t + \alpha_0$, while ‘entropy’ corresponds to the differential entropy computed as:

$$-\int_{-\pi}^{\pi} \hat{f}(\lambda) \log\left(\hat{f}(\lambda)\right) \, \mathrm{d}\lambda \, , \tag{3}$$

where $\hat{f}(\lambda)$ is an estimate of the spectral density of the data. An entropy value close to ‘0’ represents a high signal content, while a value close to ‘1’ represents a high noise content.

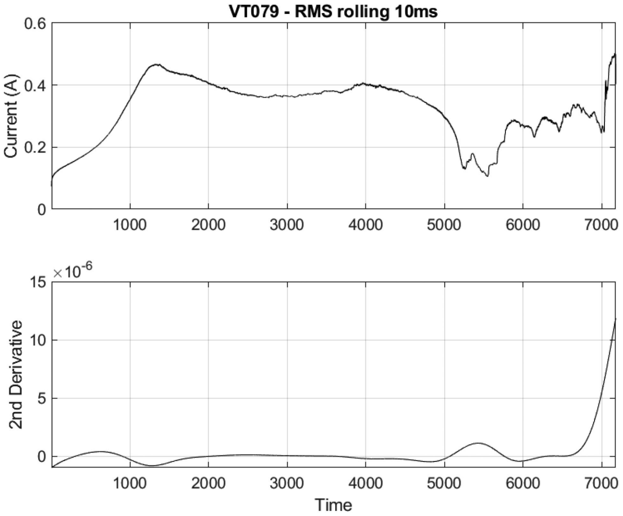

Fig. 4. LF current of test VT079. (top) RMS signal and (bottom) second derivative.

*c) Features derived from auto-correlation:* The third type of features are derived from auto-correlation analysis of $x_t$, its first difference, $\dot{x}_t = x_t - x_{t-1}$ and second difference, $\ddot{x}_t = \dot{x}_t - \dot{x}_{t-1}$. These are ‘x_acf1,’ ‘x_acf10,’ ‘diff1_acf1,’ ‘diff1_acf10,’ ‘diff2_acf1’ and ‘diff2_acf10,’ where ‘x’ indicates the original signal, ‘diff1’and ‘diff2’ represent the first and second differences $\dot{x}$ and $\ddot{x}$, ‘acf1’ corresponds to the first auto-correlation coefficient, while ‘acf10’ corresponds to the sum of the first ten squared auto-correlation coefficients. High values of ‘acf1’ and ‘acf10’ indicate the existence of complex periodical structure in the auto-correlation results, while low values indicate that the original signal or its differences are similar to white noise.

*d) Features derived from current acceleration:* The final type of time-domain features were derived from analysing the acceleration of the RMS current. As observed in Fig. 1 above, the RMS current quickly increases during Phase 1 reaching a peak, where the acceleration becomes negative and the risk of ignition is the highest. To identify this pattern, we estimate the instantaneous value of the second derivative using local polynomial regression which employs kernel smoothing [33]. To compute the second derivative, local polynomial regression fits a quadratic or a higher order polynomial of choice to each point using a neighbourhood around that point. Then the second derivative of the fitted polynomial is computed for each point [34].

Fig. 4 shows both the RMS LF current and its second derivative calculated using this method. To characterise the distribution of this signal, we take eight summary statistics as features: its mean, median, maximum, standard deviation, interquartile range, and $\{90, 95, 98\}$ percentiles. High second derivative values correspond to large increases in RMS Current, potentially causing ignition. Thus, the second derivative properties may help discriminate between the ignition and non-ignition test outcomes.

*2) Frequency-Domain Features:* Both Marxsen [2], [3] and Gomes *et al.* [4] observed that high-frequency voltage noise

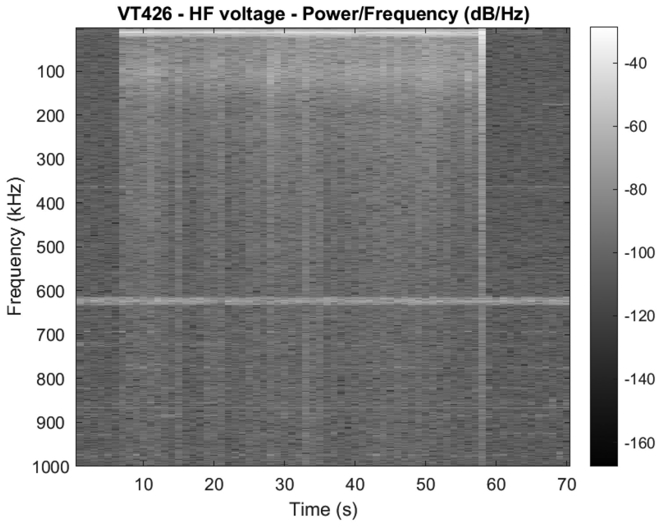

Fig. 5. Spectrum of the high-frequency voltage signal during test VT426.

correspond to a fault signature, which is caused by fast changes in the fault impedance. In normal conditions, powerline networks do not contain internal high-frequency sources; hence, they are deemed ‘quiet’. Fig. 5 illustrates the spectrum of the HF voltage signal during test VT426, whose RMS current signal was illustrated in Fig. 1. In the spectra, the 50 Hz fundamental frequency can be observed as a solid band at the top, Phase 1 occurs from 5.4 s to 19.7 s, and the Flashover corresponds to the high energy burst at 57.4 s.

We use a simple approach to characterize the signal spectra into 128 features, 32 from each one of the four raw signals. First, we compute the amplitude of one-sided Fast Fourier Transform (FFT) on the signal values from each window. The resulting spectrum is then divided into four equally-wide bands: lower $[0,\ 0.25 \times f_s]$, medium $[0.25 \times f_s,\ 0.50 \times f_s]$, high $[0.50 \times f_s,\ 0.75 \times f_s]$ and upper $[0.75 \times f_s,\ f_s]$, where $f_s$ corresponds to the sampling frequency. To characterize the distribution of the amplitude at each band, we compute its mean, median, maximum, standard deviation, interquartile range, and $\{90, 95, 98\}$ percentiles.

### *D. Classification Framework*

*1) Labelling the Data:* As described in Section III-B, we are interested in characterising the risk of ignition during Phase 1. Predicting the exact time that ignition occurs is out of the scope of this paper, as it is unknown but takes place from Phase 2 onward. Given that the rate in which the irreversible physical changes that the vegetation sample suffers is unknown, we assume a linear increase of the risk for simplicity. Let us define the start, middle and end points of Phase 1 as $t_0$, $t_{0.5}$ and $t_1$ respectively, whereby $t_0$ corresponds to the point where there is 0% risk, $t_{0.5}$ corresponds to 50% risk; and $t_1$ corresponds to 100% risk. Then, for each test where ignition develops, we label each observation between $t_0$ and $t_{0.5}$ as low-risk or ‘0,’ and between $t_{0.5}$ and $t_1$ as high-risk or ‘1’. Any other test where no ignition develops is labeled as ‘0’. Our aim is to facilitate both early detection and prevention of ignition. Note that for each test, Phase 1 duration falls between 50 s with a median of 8 s, and a mean of 10.18 s as illustrated in Fig. 6. The majority of tests reach Phase 2 within 10 s. Such a wide distribution may be due to the moisture content, the diameter of the sample, or the presence of leaves. Therefore, each test has different number of observations from each class.

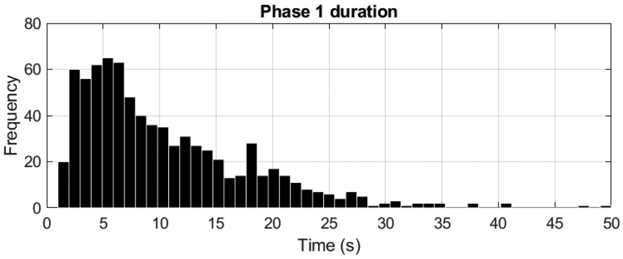

Fig. 6. Histogram of the duration of Phase 1 in seconds.

TABLE I
ACTUAL AND PREDICTED CLASS LABELS, WITH THE RF RESULT AT EACH SECOND OF PHASE 1 FOR VT386, FOR WHICH WE ACHIEVE AN ACCURACY OF 0.875

| Time (s) | $y_{i,j,k}$ | $\hat{y}_{i,j,k}$ | RF result |
|---|---|---|---|
| 1 | 0 | 0 | 0.134 |
| 2 | 0 | 0 | 0.258 |
| 3 | 0 | 0 | 0.456 |
| 4 | 0 | 1 | 0.504 |
| 5 | 1 | 1 | 0.602 |
| 6 | 1 | 1 | 0.540 |
| 7 | 1 | 1 | 0.702 |
| 8 | 1 | 1 | 0.734 |

*2) Random Forests Model:* To classify the data, we train a Random Forest (RF) model [35], a type of classification model composed of an *ensemble* of decision trees. Each independent tree is trained with a unique *bootstrap* sample of the data, that is, from a dataset of $N$ observations, we sample with replacement $N$ observations. As such, the resulting bootstrap sample has roughly 60% unique observations, and the remaining are repeats. The observations not in the bootstrap sample are known as out-of-sample (OOS) data. Because each tree is trained with different data, their response will differ for any new observation. The result from the ensemble is the class selected by the majority of the trees. For our model, we used the R implementation of the algorithm [36] with 500 decision trees.

RF models, beside being very easy to use, also provide an estimation of the importance of each feature using the OOS data, as follows. First, the error of the model on the OOS data is estimated. Then, taking one variable at the time, the values of the OOS data are randomly permuted and the error is re-estimated. The change in error determines the importance of the variable, with larger changes indicating higher importance. The results are averaged across all trees and divided by the standard deviation over the entire ensemble, resulting in a standardized measure.

*3) Model Validation:* We used 10-fold cross-validation type experiment to validate the classifier. Each one of the 1038 fault tests on the three geometries described in Section II-B was placed into one of ten subsets. That is, nine subsets are used

TABLE II
CROSS VALIDATION ACCURACY AND PERCENTAGE OF FALSE POSITIVES, FOR FAULT TESTS WITH A PHASE 1 DURATION TIME GREATER THAN $t$ SECONDS

| Fold | $t = 0$ | | $t = 4$ | | $t = 6$ | | $t = 8$ | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha_k$ | FPR | $\alpha_k$ | FPR | $\alpha_k$ | FPR | $\alpha_k$ | FPR |
| 1 | 0.870 | 0.031 | 0.888 | 0.028 | 0.899 | 0.018 | 0.908 | 0.000 |
| 2 | 0.774 | 0.060 | 0.821 | 0.065 | 0.739 | 0.063 | 0.817 | 0.081 |
| 3 | 0.850 | 0.031 | 0.837 | 0.039 | 0.882 | 0.000 | 0.988 | 0.000 |
| 4 | 0.811 | 0.039 | 0.838 | 0.048 | 0.853 | 0.039 | 0.849 | 0.036 |
| 5 | 0.860 | 0.030 | 0.909 | 0.027 | 0.918 | 0.020 | 0.979 | 0.000 |
| 6 | 0.856 | 0.029 | 0.942 | 0.021 | 0.953 | 0.027 | 0.878 | 0.094 |
| 7 | 0.828 | 0.049 | 0.882 | 0.034 | 0.883 | 0.020 | 0.849 | 0.026 |
| 8 | 0.848 | 0.043 | 0.838 | 0.014 | 0.871 | 0.007 | 0.806 | 0.016 |
| 9 | 0.827 | 0.012 | 0.842 | 0.013 | 0.816 | 0.000 | 0.903 | 0.000 |
| 10 | 0.803 | 0.036 | 0.816 | 0.049 | 0.742 | 0.035 | 0.724 | 0.036 |
| Mean | 0.833 | 0.036 | 0.861 | 0.034 | 0.856 | 0.023 | 0.870 | 0.029 |
| Std. dev. | 0.030 | 0.013 | 0.042 | 0.017 | 0.071 | 0.019 | 0.080 | 0.034 |

for training and the remaining one for testing, repeating this for each one of the subsets. The fraction of '1' outputs from each decision tree corresponds to the RF result, i.e. the proportion of decision trees predicting ignition. Using a cutoff of 0.500, we label the result as ignition or non-ignition. For each fault test, we compare the actual with the predicted labels and compute the classification accuracy for that test, as follows. Let $y_{i,j,k}$ be the actual label for the observation at time $i$ of fault test $j$ from fold $k$, $\hat{y}_{i,j,k}$ the predicted label, $\mathbf{1}(\cdot)$ be the indicator function, $T_{j,k}$ the Phase 1 duration of fault test $j$ from fold $k$, and $N_k$ the number of fault tests on each fold. Then, the test accuracy is defined as:

$$\alpha_{j,k} = \frac{1}{T_{j,k}} \sum_{i=1}^{T_{j,k}} \mathbf{1}\left(y_{i,j,k} = \hat{y}_{i,j,k}\right) \tag{4}$$

For example, Table I shows the results for the VT386 test for which we achieve an accuracy of 0.875. For this test, ignition started at some point after the end of Phase 1 at 8 s. However, we predict risk of ignition at 4 s facilitating early detection. Because of the uneven duration of Phase 1, as observed in Fig. 6, we define the accuracy for each fold using Eq (5).

$$\alpha_k = \frac{1}{N_k} \sum_{j=1}^{N_k} \alpha_{j,k} \tag{5}$$

It is worth noting that there are 263 fault tests with a Phase 1 duration lasting less than 5 s, of which some failed to ignite. To understand how our model performs on longer tests, we also calculate the accuracy for tests with Phase 1 duration at least 4 s, 6 s or 8 s.

*4) Feature Selection:* As we computed 240 features in total (112 time- and 128 frequency-domain), some may be redundant. To eliminate these redundancies, simplify the model and speed up computation, we group them based on their dissimilarity defined as $1 - |\rho|$, where $\rho$ is their Pearson correlation. For this purpose, we employ the Partition Around Medoids (PAM) clustering algorithm, which divides the features into $k$ clusters around the medoids, i.e., the observations from the cluster whose average dissimilarity to all the objects in the cluster is minimal. We use the R implementation of this algorithm [37]. To determine the value of $k$, we use the silhouette score [38] which measures how similar items are within a cluster compared with items in other clusters. A value close to '1' indicates good clustering, while a value close to '$-1$' indicates poor clustering. As such, we select the lowest value of $k$ whose difference to the maximum silhouette score value is the smallest.

Once grouped, we proceed to select one feature from each cluster and train a RF model. We select the feature combination which provides the model with the highest cross-validated accuracy. However, the possible number of feature combinations for $k$ clusters with the $i^{\text{th}}$ cluster containing $n_i$ features is $\prod_{i=1}^{k} n_i$, which can be a very large number. Therefore, we test 1000 randomly selected feature combinations. Once a good feature combination is identified and the RF model trained, we estimate the importance of each feature to determine the risk of ignition using the RF importance score.

## IV. RESULTS

### A. Analysis Using Low and High Frequency Data

Table II shows the results from applying the methodology described in Section III, using the complete set of 240 features. We see that the accuracy per fold $\alpha_k$ improves as $t$ increases. For example, we can predict ignition with an average accuracy of 0.8614 for an experiment that had a first phase duration of 4 seconds or greater. In addition to classification accuracy, we also include the proportion of false positives (FPR) in Table II, i.e. the fraction of tests falsely predicted as resulting in ignition. The proportion of false positives remain stable between 2% to 3% for all $t \in \{0, 4, 6, 8\}$.

The results of the silhouette analysis for $k$ clusters, $k \in \{2, 239\}$ gives the average silhouette score shown in Fig. 7. The maximum average silhouette score of 0.5677 is obtained with $k = 78$ clusters. However, an average silhouette score of 0.5165 is achieved using $k = 36$, which is less than half of 78 clusters. As such, we will use $k = 36$ in our computations. For the 36 cluster configuration, each cluster contains some number of features. If we denote the number of features in the $j^{\text{th}}$ cluster by $n_j$, then the sequence $\{n_j\}_{j=1}^{36}$ denotes the number of features in each cluster. Fig. 8 shows a histogram of this sequence, i.e. the number of clusters with $n$ features for $n \in \{1, \ldots, 28\}$ for $k = 36$. If we use all feature combinations in the 36 clusters we

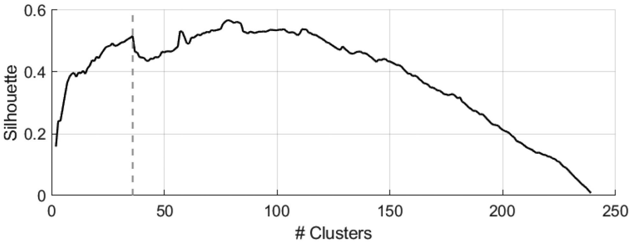

Fig. 7. The average silhouette score of a silhouette analysis using different number of clusters.

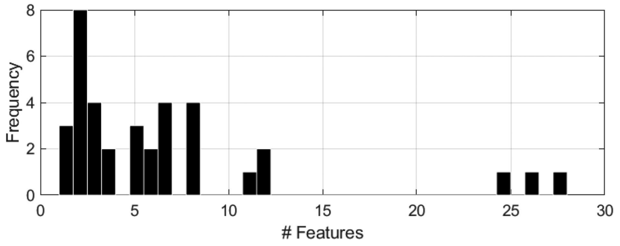

Fig. 8. Frequency of the number of features per cluster for $k = 36$.

TABLE III
SUMMARY STATISTICS OF A SERIES OF RANDOM FOREST MODELS USING 1000 FEATURE COMBINATIONS FOR $k = 36$

| Summary Statistic | ACC |
|---|---|
| Minimum | 0.810 |
| First Quartile | 0.822 |
| Median | 0.826 |
| Mean | 0.826 |
| Third Quartile | 0.829 |
| Maximum | 0.838 |

TABLE IV
LIST OF 36 FEATURES THAT YIELD THE HIGHEST 10 FOLD CV ACCURACY

| Signal | Feature Name |
|---|---|
| LF Current Raw | spike, linearity, diff2_acf1, lband_max, lband_q95, hband_mean, hband_q98 |
| LF Current RMS | linearity, curvature, x_acf1, diff1_acf1, dev2_q90 |
| HF Current Raw | diff1_acf1, diff2_acf10, hband_max |
| HF Current RMS | spike, x_acf1, diff2_acf1 |
| LF Voltage Raw | curvature, entropy, mband_max, hband_sd, hband_q98 |
| LF Voltage RMS | entropy, x_acf10 |
| HF Voltage Raw | curvature, x_acf1, diff1_acf10, mband_iqr, mband_max, uband_iqr, uband_max |
| HF Voltage RMS | trend, linearity, curvature, x_acf10 |

obtain a total number of $4.232882 \times 10^{23}$ feature combinations, therefore we only test 1000 randomly selected combinations. The performance summary of this model is given in Table III. The feature combination that yielded the highest 10 fold cross validation accuracy is given in Table IV.

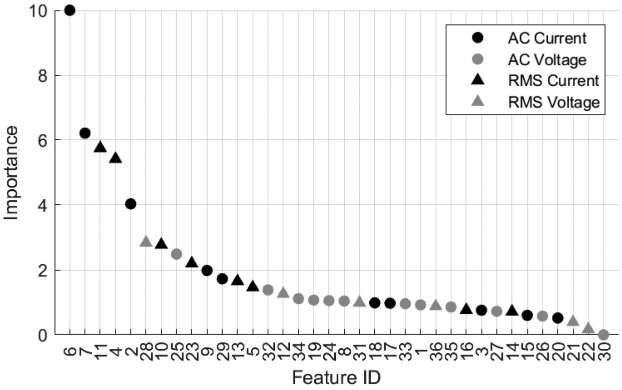

Fig. 9. The importance of features representing the 36 clusters according to the Random Forest model.

The features in Table IV that are joined by underscores have a reference to the data used in the feature in its first part and a reference to the computation of the feature in its second part. For example diff2_acf1 gives the first auto-correlation term of the twice differenced time series. If we examine the first part of these joined features, the terms x_, diff1_ and diff2_ refer to the original, once differenced and twice differenced time series. The terms lband_, mband_, hband_ and uband_ refers to the four frequency bands, lower, medium, high and upper. The term dev2_ refers to the second derivative of the LF current RMS signal. As to the second part of these joined features, the terms _acf1 and _acf10 refer to the first auto-correlation term and the sum of the first ten auto-correlation terms. The terms _q90, _q95, _q98 and _max refer to the respective percentiles and the maximum. The terms _sd, and _iqr refers to the standard deviation and IQR.

The relative importance of these 36 features according to the Random Forest model is given in Fig. 9. We see that features from current signals are generally ranked higher than those from voltage signals, with the top five having a higher importance score than the rest. These are:

1) lband_max_LF_current_Raw is the maximum of the lower frequency band of the LF raw current
2) diff2_acf1_LF_current_Raw is the first auto-correlation coefficient of the second order differences of the LF raw current
3) linearity_LF_current_RMS
4) curvature_LF_current_RMS
5) spike_LF_current_Raw

These five features are based on the LF current signals. Features such as ‘curvature’ and ‘linearity’ measure rates of growth in the current, while ‘spikiness’ indicates sudden jumps in the current. As such, ignition is mostly determined by a sudden acceleration on the rate of current growth.

### B. Analysis Using Low Frequency Data Only

One of the challenges faced in a non-experimental, real setting is accessing HF currents and voltages. Sensing is often made at the substation by SCADA or PMU systems, which tend to be located away from the fault. Due to the attenuation of HF components the signals received may not resemble those used to train the RF classifier. On the other hand, placing HF sensors at

TABLE V
CROSS VALIDATION ACCURACY AND PERCENTAGE OF FALSE POSITIVES, FOR FAULT TESTS WITH A PHASE 1 DURATION TIME GREATER THAN $t$ SECONDS USING ONLY LOW FREQUENCY CURRENT AND VOLTAGES

| Fold | $t=0$ | | $t=4$ | | $t=6$ | | $t=8$ | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha_k$ | FPR | $\alpha_k$ | FPR | $\alpha_k$ | FPR | $\alpha_k$ | FPR |
| 1 | 0.864 | 0.044 | 0.849 | 0.050 | 0.867 | 0.038 | 0.865 | 0.022 |
| 2 | 0.784 | 0.060 | 0.819 | 0.066 | 0.748 | 0.052 | 0.799 | 0.097 |
| 3 | 0.847 | 0.039 | 0.838 | 0.048 | 0.854 | 0.025 | 0.930 | 0.056 |
| 4 | 0.807 | 0.044 | 0.820 | 0.056 | 0.833 | 0.052 | 0.797 | 0.071 |
| 5 | 0.868 | 0.032 | 0.915 | 0.022 | 0.899 | 0.039 | 0.958 | 0.000 |
| 6 | 0.848 | 0.042 | 0.918 | 0.026 | 0.931 | 0.036 | 0.878 | 0.093 |
| 7 | 0.814 | 0.068 | 0.872 | 0.052 | 0.889 | 0.026 | 0.858 | 0.033 |
| 8 | 0.839 | 0.049 | 0.841 | 0.004 | 0.862 | 0.007 | 0.787 | 0.015 |
| 9 | 0.819 | 0.022 | 0.810 | 0.043 | 0.812 | 0.026 | 0.837 | 0.083 |
| 10 | 0.794 | 0.045 | 0.804 | 0.058 | 0.707 | 0.059 | 0.688 | 0.053 |
| Mean | 0.828 | 0.044 | 0.849 | 0.043 | 0.840 | 0.036 | 0.840 | 0.052 |
| Std. dev. | 0.027 | 0.012 | 0.038 | 0.018 | 0.065 | 0.014 | 0.073 | 0.030 |
| Difference in Mean | 0.005 | | 0.012 | | 0.016 | | 0.030 | |

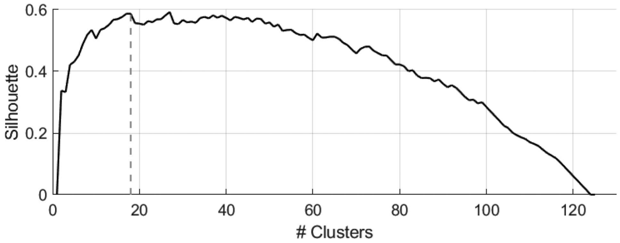

Fig. 10. The average silhouette score using only low frequency signal features.

TABLE VI
COMPARISON OF SUMMARY STATISTICS OF A SERIES OF RANDOM FOREST MODELS USING 1000 LOW FREQUENCY FEATURE COMBINATIONS FOR $k = 18$

| Summary Statistic | ACC |
|---|---|
| Minimum | 0.809 |
| First Quartile | 0.820 |
| Median | 0.823 |
| Mean | 0.823 |
| Third Quartile | 0.827 |
| Maximum | 0.837 |

closer intervals could be economically infeasible. To address this issue, we repeat the analysis from Section IV-A using only the LF data, which can be sensed at less frequent intervals. Table V gives the RF 10-fold cross validation accuracy results using all the LF features, with the bottom row presenting the difference in mean of these results and those of Table II. We observe that the maximum difference of 0.030 occurs for $t = 8$ s, whereas the minimum difference of 0.005 occurs for $t = 0$. Using the complete set of 240 features gave a maximum accuracy of 87% as seen from Table II. This was reduced to 84% when we used only low frequency features to predict ignition.

Fig. 10 shows the average silhouette score for $k \in \{2, 125\}$, with the first peak at $k = 18$, which we use from now onward. Table VI gives the equivalent of Table III using LF voltage and current features, i.e. the RF summary statistics of 1000 feature combinations using $k = 18$. We see that the difference in performance using low frequency features and $k = 36$ on all features is less that 1%. Similarly Fig. 11 shows the importance of the 18 features using the RF model. The five most important features of this cluster are: (a) lband_sd_LF_current_Raw (b) dev2_iqr_LF_current_RMS (c) spike_LF_current_Raw (d) curvature_LF_current_RMS (e) linearity_LF_current_RMS.

Comparing these 5 features with those obtained using in Section IV-A, we note that 3 of them (spike, curvature and linearity) are common. Furthermore, the features lband_sd_LF_current_Raw and lband_max_LF_current_Raw are related because we can expect the standard deviation to increase when the maximum increases. Thus, there is a strong agreement between these two sets of best 5 features.

TABLE VII
LIST OF 18 FEATURES THAT YIELD THE HIGHEST 10 FOLD CV ACCURACY

| Signal | Feature Name |
|---|---|
| LF Current Raw | lband_sd, spike, trend, linearity, mband_sd |
| LF Current RMS | dev2_iqr, curvature, linearity, spike, diff1_acf10, |
| LF Voltage Raw | e_acf10, uband_sd, mband_sd, diff2_acf10, linearity |
| LF Voltage RMS | x_acf10, entropy, curvature |

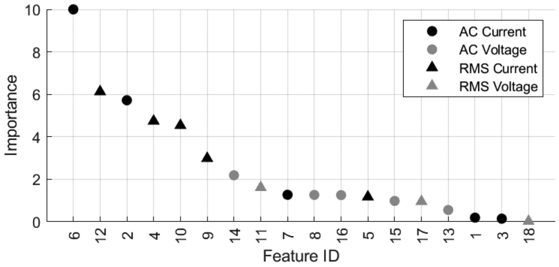

Fig. 11. The importance of features representing the 18 clusters according to the Random Forest model.

## V. Conclusion

In this paper, we proposed a methodology for prevention of wildfires, through the accurate prediction and early detection of the ignition risk resulting from HIFs. Our methodology relies on a set of over 200 simple features commonly used anomaly detection, derived from both the time- and frequency-domain analysis. The features also include a new subset based on the instantaneous estimation of the second derivative of the RMS current, using local polynomial regression. We tested our methodology on the large PBSP dataset, which is stored in the proprietary `.pnrf` file format, widely used on data acquisition tasks but uncommon for the broader Machine Learning community. As such, we provided the `pnrfr` package [16], a wrapper of the PNRF Toolkit Reader software by HBM for the R programming language; hence, it is limited to the Windows Operating System at the moment. However, for the researchers working with `.pnrf` datasets, this package gives them access to a broad array of state-of-the-art methods for classifying complex data available in R. The results from our tests demonstrate that we are able to: (a) detect the probability of ignition with high accuracy and well before its occurrence; and (b) use only a smaller subset of 36 features, reducing the computational burden for real time implementation. We identified that the features associated with current had the largest influence in determining the risk of ignition and further narrowed down the features of importance. Finally, we compared the results using only LF data, as HF data may be difficult to capture in the field. We observed a statistically significantly reduction in average accuracy of 1.5%.

There are limitations to our research. For example, the data employed was collected in a laboratory setting, where many conditions are tightly controlled. As such, we have not explored many of the intricacies that an on-the-field, real-time implementation would require, including the efficient and cost-effective collection of HF data. Moreover, we have no evidence that the selected features would describe the fault signatures at voltages other than 12.7 kV phase-to-earth and 22 kV phase-to-phase, nor whether the signatures would be significantly different for a different set of vegetation samples. Despite this, our methodology readily generalises to other experimental conditions whenever more comprehensive data becomes available. Therefore, these issues are left for further research.

One of the objectives of the Powerline Bushfire Safety Program (PBSP) was to identify the worst species for fire starts from powerline faults and understand their ignition processes. As such, the PBSP dataset comprised experiments using different types of vegetation native to the Australian forests and grasslands. As our focus was estimating the risk of ignition, our experiments disregarded the information on the vegetation species. Another avenue for further research is to examine the effects that the species have on ignition, and possibly predicting which one has produced the fault.

## Acknowledgment

The authors acknowledge the Powerline Bushfire Safety Program for funding the large scale experimental work that derived in the data used for this research.

## References

[1] C. Miller *et al.*, “Electrically caused wildfires in Victoria, Australia are over-represented when fire danger is elevated,” *Landscape Urban Planning*, vol. 167, pp. 267–274, 2017.

[2] T. Marxsen, “Vegetation conduction ignition test report – final,” Powerline Bushfire Safety Program, Dept. Econ. Develop., Jobs, Transport Resour., Tech. Rep., 2015.

[3] T. Marxsen, “New technology to cut Victoria’s powerline fire risk,” presented at the Proc. Arboriculture Aust. Nat. Conf., 2016.

[4] D. P. Gomes, C. Ozansoy, A. Ulhaq, and J. C. de Melo Vieira Júnior, “The effectiveness of different sampling rates in vegetation high-impedance fault classification,” *Electric Power Syst. Res.*, vol. 174, 2019, Art. no. 105872.

[5] A. Ghaderi, H. L. Ginn, and H. A. Mohammadpour, “High impedance fault detection: A review,” *Electric Power Syst. Res.*, vol. 143, pp. 376–388, 2017.

[6] S. Jazebi, F. de Leon, and A. Nelson, “Review of wildfire management techniques—Part I: Causes, prevention, detection, suppression, and data analytics,” *IEEE Trans. Power Del.*, vol. 35, no. 1, pp. 430–439, Feb. 2020.

[7] S. Jazebi, F. de Leon, and A. Nelson, “Review of wildfire management techniques—Part II: Urgent call for investment in research and development of preventative solutions,” *IEEE Trans. Power Del.*, vol. 35, no. 1, pp. 440–450, Feb. 2020.

[8] W. Zhang, Y. Jing, and X. Xiao, “Model-based general arcing fault detection in medium-voltage distribution lines,” *IEEE Trans. Power Del.*, vol. 31, no. 5, pp. 2231–2241, Oct. 2016.

[9] A. Mukherjee, A. Routray, and A. K. Samanta, “Method for online detection of arcing in low-voltage distribution systems,” *IEEE Trans. Power Del.*, vol. 32, no. 3, pp. 1244–1252, Jun. 2017.

[10] S. Guggenmoos, “Effects of tree mortality on power line security,” *J. Arboriculture*, vol. 29, no. 4, pp. 181–196, 2003.

[11] N. Bahador, H. R. Matinfar, and F. Namdari, “A framework for wide-area monitoring of tree-related high impedance faults in medium-voltage networks,” *J. Elect. Eng. Technol.*, vol. 13, no. 1, pp. 1–10, 2018.

[12] C. L. Benner, R. A. Peterson, and B. D. Russell, “Application of DFA technology for improved reliability and operations,” in *Proc. IEEE Rural Electric Power Conf.*, 2017, pp. 44–51.

[13] D. P. S. Gomes, C. Ozansoy, and A. Ulhaq, “High-sensitivity vegetation high-impedance fault detection based on signal’s high-frequency contents,” *IEEE Trans. Power Del.*, vol. 33, no. 3, pp. 1398–1407, Jun. 2018.

[14] D. P. S. Gomes, C. Ozansoy, and A. Ulhaq, “Vegetation high-impedance faults’ high-frequency signatures via sparse coding,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 5233–5242, 2019.

[15] R. A. Broome and W. T. Smith, “The definite health risks from cutting power outweigh possible bushfire prevention benefits,” *Med. J. Aust.*, vol. 197, no. 8, pp. 440–441, 2012.

[16] N. Antharama, *pnrfr: Read pnrf files*, 2019. [Online]. Available: https://github.com/nandinisa/pnrfr/

[17] B. Teague, R. McLeod, and S. Pascoe, “Final Report,” 2009 Victorian Bushfires Royal Commision, Tech. Rep., 2010. [Online]. Available: http://royalcommission.vic.gov.au/Commission-Reports/Final-Report.html

[18] Energy Safe Victoria, “Managing trees near powerlines,” 2020. [Online]. Available: https://esv.vic.gov.au/technical-information/electrical-installations-and-infrastructure/managing-trees-near-powerlines/

[19] J. V. Di Giulio, “Power line politics and bushfire safety: Analysis of the policy implications from the Victorian Bushfires Royal Commission,” 2010. [Online]. Available: https://www.researchgate.net/publication/311769119_Power_Line_Politics_and_Bushfire_Safety_Analysis_of_the_policy_implications_from_the_Victorian_Bushfires_Royal_Commission

[20] R. Oloruntoba, “Plans never go according to plan: An empirical analysis of challenges to plans during the 2009 Victoria bushfires,” *Technological Forecasting Social Change*, vol. 80, no. 9, pp. 1674–1702, 2013.

[21] M. Eburn and S. Dovers, “Learning lessons from disasters: Alternatives to Royal Commissions and other quasi-judicial inquiries,” *Aust. J. Public Admin.*, vol. 74, no. 4, pp. 495–508, 2015.

[22] M. Williamson *et al.*, “Solar PV and energy storage: A solution to powerline-related bushfires,” *Energy News*, vol. 34, no. 2, pp. 12–16, 2016.

[23] R. Roozbahani, C. Huston, S. Dunstall, B. Abbasi, A. Ernst, and S. Schreider, “Minimizing bushfire risk through optimal powerline assets replacement and improvement,” in *Proc. 21st Int. Congr. Modelling Simul.*, 2015, pp. 1834–1840.

[24] J. Whittaker, J. Handmer, and D. Karoly, "After 'Black Saturday': Adapting to bushfires in a changing climate," in *Natural Disasters and Adaptation to Climate Change*, S. Boulter, J. Palutikof, D. J. Karoly, and D. Guitart, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2013, pp. 75–86.

[25] B. Stewart, "Design and operation of high voltage R&D test facility for bushfire mitigation technologies," in *Proc. Chemeca: Chem. Eng.-Regeneration, Recovery Reinvention*, Melbourne, Vic.: Engineers Australia, 2016, pp. 1007–1018.

[26] P. Montero-Manso, G. Athanasopoulos, R. J. Hyndman, and T. S. Talagala, "FFORMA: Feature-based forecast model averaging," *Int. J. Forecasting*, vol. 36, no. 1, pp. 86–92, 2020.

[27] B. D. Fulcher and N. S. Jones, "Highly comparative feature-based time-series classification," *IEEE Trans. Knowl. Data Engi.*, vol. 26, no. 12, pp. 3026–3037, Dec. 2014.

[28] T. Räsänen and M. Kolehmainen, "Feature-based clustering for electricity use time series data," in *Proc. Int. Conf. Adaptive Natural Comput. Algorithms*, 2009, pp. 401–412.

[29] R. J. Alcock *et al.*, "Time-series similarity queries employing a feature-based approach," in *Proc. 7th Hellenic Conf. Informat.*, 1999, pp. 27–29.

[30] P. Talagala, R. Hyndman, K. Smith-Miles, S. Kandanaarachchi, and M. A. Muñoz, "Anomaly detection in streaming nonstationary temporal data," *J. Comput. Graphical Statist.*, vol. 29, no. 1, pp. 13–27, 2020.

[31] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on LOESS," *J. Official Statist.*, vol. 6, no. 1, pp. 3–73, 1990.

[32] R. Hyndman *et al.*, *tsfeatures: Time Ser. Feature Extraction*, 2019, R package version 1.0.1. [Online]. Available: https://CRAN.R-project.org/package=tsfeatures

[33] M. Wand, *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2015, R package version 2.23-15. [Online]. Available: https://CRAN.R-project.org/package=KernSmooth

[34] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman & Hall, 1994.

[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[36] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[37] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, *Cluster: Cluster Anal. Basics and Extensions*, 2018, R package version 2.1.0.

[38] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. Supplement C, pp. 53–65, 1987.

**Sevvandi Kandanaarachchi** received the B.Sc.Eng. in 2002 from the University of Moratuwa, Sri Lanka. She completed her Ph.D. degree in mathematics in 2011 from Monash University, Australia. In 2015, she completed a Graduate Certificate in data mining and applications from Stanford University, United States. Her research interests include anomaly and event detection, dimension reduction and algorithm evaluation. She is a Lecturer at RMIT University and a Research Affiliate at Monash University.

**Nandini Anantharama** is currently a Ph.D. Candidate at Monash University, Australia. Her research interests include the applications of machine learning towards building exploratory, predictive and decision making models. Her research focus is mining electronic health records. She is also interested in the use of deep neural nets in the modeling of free-form text data such as clinical notes.

**Mario A. Muñoz** received the B.Eng. and M.Eng. degrees in electronics engineering from Universidad del Valle, Colombia, in 2005 and 2008 respectively, and the Ph.D. degree in engineering from The University of Melbourne, Australia, in 2014. Currently he is a Research Fellow (Level B) at the School of Mathematics and Statistics, The University of Melbourne. He has published over 40 papers.

His research interests focus on the application of optimization, computational intelligence, signal processing, data analysis, and machine learning methods to ill-defined science, engineering and medicine problems.