

Article

# Entropy-Based Feature Extraction for Electromagnetic Discharges Classification in High-Voltage Power Generation

Imene Mitiche <sup>1,\*</sup>, Gordon Morison <sup>1</sup>, Alan Nesbitt <sup>1</sup>, Brian G. Stewart <sup>2</sup> and Philip Boreham <sup>3</sup>

<sup>1</sup> Department of Engineering, Glasgow Caledonian University, 70 Cowcaddens Rd, Glasgow G4 0BA, UK; gordon.morison@gcu.ac.uk (G.M.); a.nesbitt@gcu.ac.uk (A.N.)

<sup>2</sup> Institute of Energy and Environment, University of Strathclyde, 204 George St, Glasgow G1 1XW, UK; brian.stewart.100@strath.ac.uk

<sup>3</sup> Innovation Centre for Online Systems, 7 Townsend Business Park, Bere Regis BH20 7LA, UK; pboreham@doble.com

\* Correspondence: imene.mitiche@gcu.ac.uk; Tel.: +44-(0)-141-331-3717

Received: 22 May 2018; Accepted: 20 July 2018; Published: 25 July 2018



**Abstract:** This work exploits four entropy measures known as Sample, Permutation, Weighted Permutation, and Dispersion Entropy to extract relevant information from Electromagnetic Interference (EMI) discharge signals that are useful in fault diagnosis of High-Voltage (HV) equipment. Multi-class classification algorithms are used to classify or distinguish between various discharge sources such as Partial Discharges (PD), Exciter, Arcing, micro Sparking and Random Noise. The signals were measured and recorded on different sites followed by EMI expert's data analysis in order to identify and label the discharge source type contained within the signal. The classification was performed both within each site and across all sites. The system performs well for both cases with extremely high classification accuracy within site. This work demonstrates the ability to extract relevant entropy-based features from EMI discharge sources from time-resolved signals requiring minimal computation making the system ideal for a potential application to online condition monitoring based on EMI.

**Keywords:** EMI measurement; partial discharge; entropy; classification; experts system; EMI discharge sources

## 1. Introduction

The High-Voltage (HV) power supply industry involves the use of generators, motors, transformers, transmission lines and cables that jointly contribute to generate electrical power. The effective operation of these assets is important as any failure in one part of the system may have significant and often drastic consequences in terms of loss of production, induced costs, safety and down-time. Early detection or prediction of fault occurrence along with fault source identification would potentially avoid such dramatic consequences as appropriate risk mitigation measures could be put in place. Electrical power assets such as generators, motors and transformers are susceptible to electrical insulation problems [1] which can be manifested in the form of faults such as Partial Discharge (PD) and Arcing. PD is a sign of insulation degradation which is considered harmful for an asset, as once present it becomes a further source of accelerated insulation degradation [2]. As such, PD monitoring is seen as a significant tool for electrical insulation condition assessment [3]. PD can be assigned several different PD categories such as Surface Discharge, and Internal Discharge [4]. An arcing fault is also an electrical discharge through an insulating medium [5]. An arc occurs when

electric current is transmitted through a gaseous or liquid environment. These latter two faults commonly arise in power transformers [6]. Early detection of these faults helps to take decisions on an asset's maintenance which reduces the high cost of purchasing new replacement equipment as well as avoiding potential statutory fines and civil complaints [4]. Identifying PD faults may be achieved from appropriate PD captured data using pattern recognition and classification techniques along with appropriate feature extraction methods [7]. This approach for PD classification has been investigated in the literature by many researchers within power systems [8,9]. In [10] the authors extracted image-oriented features from phase resolved plots of PD signals and employed different classifiers for pattern recognition, e.g., Fuzzy k-Nearest Neighbour Classifier (FkNNC), Back-Propagation Neural Network (BPNN) and Support Vector Machine (SVM). Classification of multiple PD sources (Void, Air-Corona and Oil-Corona) was achieved in [11] using statistical quantities, such as mean deviation, quartile deviation etc., calculated on the Phase Resolved PD (PRPD) patterns and classified with a Probabilistic Neural Network (PNN). A comparison between PNN and SVM classification of PD sources in [12] revealed that SVM slightly outperforms PNN in terms of classification accuracy. From the research, PD analysis and classification revealed that different discharge sources have a unique fingerprint that can be quantified with the help of feature extraction techniques.

For many years, the Electro-Magnetic Interference (EMI) method has been successfully applied to diagnose generator and transformer faults (e.g., [13,14]). This method differs from other fault diagnosis methods in that EMI looks at the frequency and time-domain measurements of captured signals using The Comite International Special des Perturbations Radioelectriques (CISPR) 16 Standard approach. In CISPR 16, different bandpass filters are used to measure the signal energy over different frequency ranges. The presence of frequency bands in a CISPR 16 spectrum is indicative of different types of generator faults, with the most important fault frequency bands existing well below 100 MHz (see [13,14]). Once a frequency band and an associated frequency of interest has been selected in an EMI spectrum, the audio time-domain signal at the selected EMI frequency is captured and reviewed by an EMI expert for "expert" fault classification, e.g., Arcing, Corona, Exciter, Data Modulation, PD etc. [15]. Currently there is a limited amount of work investigating the use of intelligent pattern recognition or signal fault classification methods for time-domain EMI sources within power system equipment. The aim of this work is to assess the potential for the use of Entropy-based methods in EMI-based condition monitoring. To this end, this paper applies four simple and computationally low feature extraction methods based on entropy along with Multi-Class SVM (MCSVM) and Random Forests (RF) classifiers with the aim of separating and classifying multiple EMI time-domain signals. The measurements investigated in this paper were collected from various HV power stations which were identified to contain EMI faults based on EMI analysis from field experts. More details regarding the data and classification are discussed later in this paper.

The paper is organised as follows. Section 2 discusses the EMI method of monitoring and capturing PD signals in more detail. Section 3 introduces the feature extraction and classification techniques employed. Section 4 describes the data measurement process and classification experimental set-up. Section 5 presents the results from measured signals; evaluation of the performance of the classification approach in relation to the segmented size of time domain data collected is also considered. The last section draws conclusions on the work and provides suggestions on future work.

## 2. EMI Monitoring

An EMI measurement of power equipment is carried out as follows. Using a High-Frequency-Current-Transformer (HFCT), signals are measured on the neutral earth cable of, for example a generator or transformer. The signals are usually recorded with a PD Surveyor 200 (PDS200) instrument that functions under the CISPR16 standard [16] for EMI filter types. CISPR 16 specifies the use of a quasi-peak detector implemented with different filter bandwidths over the range of frequencies to be measured. For example, the B filter operates over 150 kHz–30 MHz, with a 6 dB bandwidth of 9 kHz, the C filter from 30–300 MHz with a 6dB bandwidth of 120 kHz. The important feature of PDS200 is

it acts as a radio receiver with the ability to capture and analyse Radio Frequency Interference (RFI) as well as EMI radiations in the lower frequency range [50 kHz–1 GHz]. This provides a frequency spectrum of unique signature for each fault or condition [13,14,17]. The PDS200 looks at signals based on the CISPR16 bandwidths (e.g., 9 kHz, 120 kHz) and moves through the frequency spectrum making appropriate power filtered response measurements at each frequency. Quasi-peak measurements are made over a 1 s time-period at each selected frequency. For time domain signals the instrument permits any frequency between 150 kHz to 100 MHz to be selected e.g., where the maximum envelope energy exists. It then makes the slower demodulated response measurement which is sampled at 24 kHz. EMI time-domain signals can also be retrieved at the frequencies of interest for audio examination by EMI experts in order to determine or classify the nature of the fault [15]. This ability is based on a wealth of previous experiences of audio fault assessment and forensic confirmation. The limitation of this method is that it relies completely on the availability of an expert to ascertain the fault diagnosis. However, an experienced expert with a modest training is capable of identifying many of the common faults related to EMI signals and the experts involved in this work have extensive accumulated experience of fault diagnosis and forensic confirmation. This work attempts to capture this expert knowledge and to use this as a foundation for initial training of fault recognition algorithms and automated classification for EMI time-domain signals. To achieve this, a method of feature extraction and classification requires to be implemented, which is presented in the following section.

### 3. Description of Employed Algorithms

The experiment design is summarised in the flow diagram in Figure 1. In this work, denoising approach is implemented, on the EMI time signals, as preprocessing. It is important to apply this step to noisy signals only as denoising of a noise free signal may result in distortion which will subsequently destroy the important signal information. Thus, Peak to Average Power Ratio (PAPR) is used as a decision metric to whether the signal should be denoised.

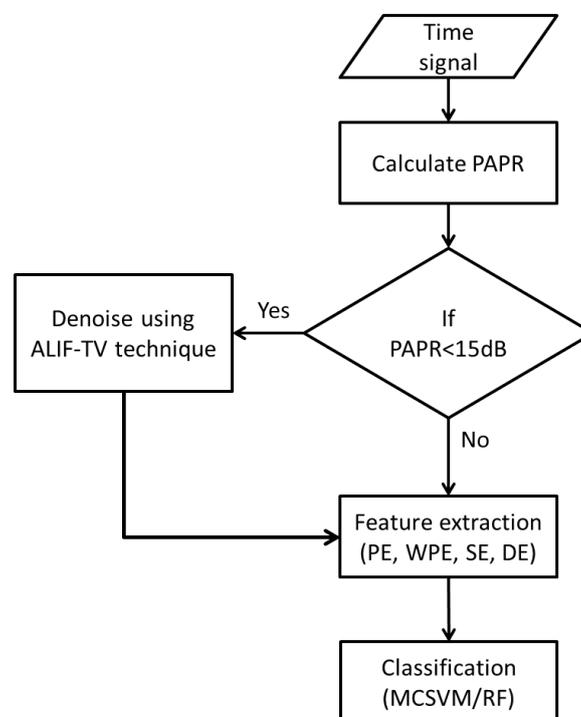


Figure 1. Flow diagram of the overall proposed algorithm.

The PAPR, defined in Equation (1), is calculated for each time signal and compared against a threshold of 15 dB. The choice of this threshold is based on evaluation of the whole experiment using different thresholds in the range of [10–20 dB] with a step of one. The main reasons of selecting this range of PAPR are under the assumption of the following. The discharge impulses such as PD have a much higher peak amplitude than noise amplitude. Therefore, a high PAPR indicates that the signal contains little noise. The common noise found in EMI measurement is similar to the one observed in telecommunication system, which typically has a PAPR that varies between [10–20 dB] [18].

$$\text{PAPR} = 10\log_{10}\left(\frac{|x_{peak}|^2}{x_{rms}^2}\right) \quad (1)$$

Relevant information is then extracted from the denoised or noise-free signals by means of Permutation Entropy (PE), Weighted Permutation Entropy (WPE), Sample Entropy (SE) and Dispersion Entropy (DE). The four entropies are embedded into a feature vector per signal instance, which is implemented in MCSVM and RF classifiers. The employed denoising, feature extraction and classification algorithms are explained in detail as follows.

### 3.1. Signal Denoising

The employed algorithm is called Adaptive Local Iterative Filtering combined with Total Variation (ALIF-TV) and was proposed in [19] to effectively mitigate noise in field measured EMI signals. The main idea of this method is to first decompose the signal, using ALIF, into its different frequency components also called Intrinsic Mode Function (IMF), then employ TV thresholding to reduce noise towards zero. The denoised signal is reconstructed as the sum of the components after thresholding. First, we denote scalars by lower case, vectors by bold lower case and matrices by bold upper case. The mathematical framework of this algorithm is described as follows.

Let the time series signal,  $\mathbf{y}(n); n = 1, \dots, N$ , be decomposed into  $K$  number of IMFs plus a residual trend  $\mathbf{r}(n)$  as:

$$\mathbf{y}(n) = \mathbf{imf}_1(n) + \mathbf{imf}_2(n) + \dots + \mathbf{imf}_K(n) + \mathbf{r}(n). \quad (2)$$

The IMFs are obtained through filtering as:

$$\mathbf{y}_{j+1}(n) = \mathbf{y}_j(n) - \sum_{-l_j(n)}^{l_j(n)} \mathbf{h}_j(n, v) \cdot \mathbf{y}_j(n + v) \quad (3)$$

where  $\mathbf{h}_j(n, v) \in \mathbb{R}, n \in [-l_j(n), l_j(n)]$  are low pass filter coefficients at point  $n$  with length of  $2l_j(n) + 1$ . Equation (3) is iterated until it converges to obtain the IMFs, this is known as inner iteration. Outer iteration is performed and converges when the residual trend is obtained. This is delineated when the number of points in  $\mathbf{y}_j(n)$  is continuously reduced while the number of iterations increases. The ALIF algorithm is embedded with TV to obtain the IMFs of a noisy signal  $\mathbf{y}(n)$  by the function

$$A(\mathbf{y}(n)) = \begin{bmatrix} \mathbf{imf}_1(n) \\ \mathbf{imf}_2(n) \\ \vdots \\ \mathbf{imf}_K(n) \end{bmatrix} = \mathbf{W}. \quad (4)$$

Each IMF in  $\mathbf{W}$  is denoised using TV thresholding. Here the thresholding is not applied to the residual. The denoised signal is reconstructed by summation of the denoised IMFs plus the residual. The TV of a signal  $\mathbf{y}(n)$  is defined as follows:

$$\text{TV}(\mathbf{y}(n)) := \|\mathbf{D}\mathbf{y}(n)\|_1 \quad (5)$$

where  $\|\cdot\|_1$  is the  $l_1$ -norm and  $\mathbf{D}$  is the first order difference matrix defined as:

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}. \quad (6)$$

The ALIF-TV minimises the noise in the IMFs ( $\hat{\mathbf{W}}$ ) by solving the following non-convex optimisation problem:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \{F(\mathbf{W}) = \frac{1}{2} \|A(\mathbf{y}(n)) - \mathbf{W}\|_2^2 + \sum_j \lambda_j \phi(\mathbf{W}_j; \alpha_j) + \beta \|\mathbf{DB}(\mathbf{W})\|_1\} \quad (7)$$

where  $\|\cdot\|_2^2$  is the  $l_2$ -norm,  $B(\mathbf{W})$  is the reconstruction function defined as follows:

$$B(\mathbf{W}) = \left( \sum_{j=1}^K \mathbf{W}_j \right) + \mathbf{r}(n) \quad (8)$$

$\phi$  is a penalty function [20] with parameter  $\alpha_j = \frac{1}{\lambda_j}$ , and the regularisation parameters  $\beta$  and  $\lambda$  are defined as follows.

$$\beta = (1 - \eta) \sqrt{N} \frac{\sigma}{4} \quad (9)$$

$$\lambda = (2.5\eta\sigma) / \sqrt{2^j} \quad (10)$$

Here,  $\eta = 0.95$  is selected according to [20] to balance the weight between the TV and IMFs in the optimisation problem. The variance of the noise  $\sigma$  is estimated, using the popular Donoho Median Absolute Deviation (MAD) [21], in the following expression:

$$\sigma = \frac{MAD(\mathbf{y}(n))}{0.6745}. \quad (11)$$

The non-convex optimisation problem is solved by means of the split augmented Lagrangian shrinkage algorithm [22]. Variable splitting is a straightforward approach that involves the creation of a new variable, here  $\mathbf{U}$ , which serves as the argument of  $f_2$  under the constraint that  $g(\mathbf{W}) = \mathbf{U}$ . This results in a constrained problem of Equation (7) as:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \{f_1(\mathbf{W}) + f_2(\mathbf{U})\} \quad (12)$$

where  $\mathbf{U} = \mathbf{W}$  initially, and

$$f_1(\mathbf{W}) = \frac{1}{2} \|A(\mathbf{y}(n)) - \mathbf{W}\|_2^2 + \sum_j \lambda_j \phi(\mathbf{W}_j; \alpha_j) \quad (13)$$

$$f_2(\mathbf{U}) = \beta \|\mathbf{DB}(\mathbf{W})\|_1. \quad (14)$$

The augmented Lagrangian is then implemented as follows:

$$L(\mathbf{W}, \mathbf{U}, \mu) = f_1(\mathbf{W}) + f_2(\mathbf{U}) + \frac{\mu}{2} \|\mathbf{U} - \mathbf{W} - \mathbf{V}\|_2^2 \quad (15)$$

where  $\mu$  is a step-size parameter set to 1 according to [20], and  $\mathbf{V} = 0$  in the initial iteration. Hence, Equation (12) is solved in three main steps of sub-problems (Equations (16) to (18)) iteratively:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \left\{ f_1(\mathbf{W}) + \frac{\mu}{2} \|\mathbf{U} - \mathbf{W} - \mathbf{V}\|_2^2 \right\} \tag{16}$$

$$\mathbf{V} = \underset{\mathbf{V}}{\operatorname{argmin}} \left\{ f_2(\mathbf{V}) + \frac{\mu}{2} \|\mathbf{U} - \mathbf{W} - \mathbf{V}\|_2^2 \right\} \tag{17}$$

$$\mathbf{V} = \mathbf{V} - (\mathbf{U} - \mathbf{W}). \tag{18}$$

Please note that the functions  $f_1$  and  $f_2$ , and the sub-problems in Equations (16) and (17) are strictly convex. When convergence of the iterative algorithm is achieved, using the theory in [23], this solves Equation (7) and the denoised signal reconstruction is obtained by:

$$\hat{\mathbf{y}}(n) = B(\hat{\mathbf{W}}). \tag{19}$$

### 3.2. Entropy Measures

#### 3.2.1. Permutation Entropy (PE)

PE was introduced by Bandt and Pompe [24] to assess the complexity in time series data based on the comparison of successive adjacent values which are mapped to ordinal patterns. The advantage of ordinal patterns helps in providing more resilience to low frequency artefacts. This makes it suitable for measuring real-world, noisy and chaotic time series signals [25]. PE is derived from Shannon’s entropy which is a measure of the level of information contained in a data set [26]. Bandt and Pompe combined the entropy concept with symbolic dynamics to create a simple, fast to compute, robust and stable measure of regularity in short time series and to overcome classical entropy method limitations, including the requirement of long data sets and high computational cost [27]. The PE algorithm is described in [28] as follows. Based on a given time series  $\{x(1), x(2), \dots, x(N)\}$ , vectors  $\mathbf{x}(j); j = 1, 2, \dots, N - (m - 1)$  are constructed as:

$$\mathbf{x}(j) = \{x(j), x(j + 1), \dots, x(j + (m - 1))\}. \tag{20}$$

The sequence in Equation (20) is arranged to provide components in increasing order as follows:

$$\{x(j + (i_1 - 1)) \leq x(j + (i_2 - 1)) \leq \dots \leq x(j + (i_m - 1))\}. \tag{21}$$

If two successive components are equal i.e.,  $x(j + (i_1)) = x(j + (i_2))$  for  $i_1 \leq i_2$ , then their positions can be rearranged to  $x(j + (i_1)) \leq x(j + (i_2))$ . Next a different symbol series  $\mathbf{s}(l)$  is calculated for each time series  $\mathbf{x}(i)$  as:

$$\mathbf{s}(l) = (i_1, i_2, \dots, i_m) \tag{22}$$

where  $l = 1, 2, \dots, k$  ( $k = m!$ ). That is, there will be  $m!$  different symbol series or permutations  $\pi_n$ . The probability  $p(\pi_n^m)$  of each symbol sequence  $\mathbf{s}(l)$  is then calculated mathematically as:

$$p(\pi_n^m) = \frac{\sum_{j \leq N}^1 u : \operatorname{type}(u) = \pi_n(\mathbf{x}_n^m)}{\sum_{j \leq N}^1 u : \operatorname{type}(u) \in \Pi(\mathbf{x}_n^m)}. \tag{23}$$

The mapping of the symbol  $\mathbf{s}(l)$  to the ordinal pattern is denoted as  $\operatorname{type}(\cdot)$  and the  $m!$  symbols  $\{\pi_n^m\}_n^{m!}$  are denoted as  $\Pi$ . The indicator function  $1_{\mathbf{a}}(u)$  of a set  $\mathbf{a}$  is defined as:

$$1_{\mathbf{a}}(u) = \begin{cases} 1 & \text{if } u \in \mathbf{a} \\ 0 & \text{if } u \notin \mathbf{a} \end{cases}$$

Finally the value of PE can be estimated using the formula:

$$PE = - \sum_{n=1}^{m!} p(\pi_n^m) \times \log(p(\pi_n^m)). \tag{24}$$

PE should be suitable to quantify the characteristics contained in EMI signals since they are by nature non-stationary and complex to analyse.

### 3.2.2. Weighted Permutation Entropy (WPE)

WPE is derived from PE except that in WPE patterns with same amplitude variations are grouped within the same pattern [29]. Given the time series data  $\{x(1), x(2), \dots, x(N)\}$  with length  $N$ , the time data is mapped to a space of  $m$  dimension and a delay  $\tau$  in order to obtain the vector  $\mathbf{x}(j)^{m,\tau} = \{x(j), x(j + \tau), \dots, x(j + (m - 1)\tau)\}; j = 1, 2, \dots, N - (m - 1)\tau$ .  $m!$  Permutation patterns  $\pi_i$  are created with an embedded dimension of  $m$ . Each vector  $\mathbf{x}(j)$  is compared to a permutation pattern  $\pi_i$  in the  $m$  dimensional space and is weighted by a weight  $w_j$  which is obtained from the variance of each neighbour's vector  $\mathbf{x}(j)^{m,\tau}$  as:

$$w_j = \frac{1}{m} \sum_{k=1}^m (x(j + (k - 1)\tau) - \overline{\mathbf{x}(j)^{m,\tau}})^2 \tag{25}$$

where  $\overline{\mathbf{x}(j)^{m,\tau}}$  is the mean of  $\mathbf{x}(j)$  which is defined as:

$$\overline{\mathbf{x}(j)^{m,\tau}} = \frac{1}{m} \sum_{k=1}^m x(j + (k - 1)\tau). \tag{26}$$

The weighted probabilities of occurrence for each pattern group  $\pi_i^{m,\tau}$  are estimated as:

$$p_w(\pi_i^{m,\tau}) = \frac{\sum_{j \leq N} 1_{u: type(u) = \pi_i}(\mathbf{x}(j)^{m,\tau}) \times w_j}{\sum_{j \leq N} 1_{u: type(u) \in \Pi}(\mathbf{x}(j)^{m,\tau}) \times w_j}. \tag{27}$$

WPE can then be estimated based on Shannon Entropy [26] as:

$$WPE = - \sum_{i: \pi_i^{m,\tau} \in \Pi} p_w(\pi_i^{m,\tau}) \times \log(p_w(\pi_i^{m,\tau})) \tag{28}$$

It was demonstrated in [30] that WPE is effective in the analysis of non-linear time series, which is a motivation to exploit it in this paper as a feature extraction technique for discharge sources since they are by nature non-stationary. Bandt and Pompe [24] suggest embedded dimension  $m$  between 3 and 7. A high value of  $m$  provides more patterns which increases memory and computation. Since an aim in this work is to employ low computationally complex methods that could potentially be implemented in an instrument, the minimum  $m = 3$  is chosen for both PE and WPE.

### 3.2.3. Sample Entropy (SE)

SE is an entropy-based complexity measure for time series data [31] which directly measures the degree of randomness and inversely measures the degree of orderliness. The overall idea in SE is to determine the conditional probability that  $m$  consecutive data points, which are similar within a range or tolerance  $r$ , will preserve this similarity if the next data point is added, given that self-matches are ignored [32]. The steps for SE calculation are described as follows.

For the input time series  $\{x(1), x(2), \dots, x(N)\}$  of length  $N$ , vectors  $(\mathbf{z}_m(j)); j = 1, \dots, N - m$  are constructed such that:

$$\mathbf{z}_m(j) = \mathbf{x}(j + k); \quad 0 \leq k \leq m - 1. \tag{29}$$

The maximum difference between the magnitude of two vectors determines the distance between them, and this distance is compared against a tolerance  $r$  providing a proportion of two vectors,  $\mathbf{x}_m(j)$  and  $\mathbf{x}_m(i), i \neq j$  with length  $m$ , is defined as

$$s_j^m = \frac{1}{N - m - 1} \sum_{i=1, i \neq j}^{N-m} \Theta(r - \|\mathbf{z}_m(j) - \mathbf{x}_m(i)\|). \tag{30}$$

On the other hand, the proportion of two vectors,  $\mathbf{x}_m(j)$  and  $\mathbf{x}_m(i), i \neq j$  with length  $m + 1$ , is defined as:

$$s_j^{m+1} = \frac{1}{N - m - 1} \sum_{i=1, i \neq j}^{N-m} \Theta(r - \|\mathbf{z}_{m+1}(j) - \mathbf{x}_{m+1}(i)\|) \tag{31}$$

where

$$\Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{32}$$

Next, the grouping average is calculated as:

$$\mathbf{u}^m = \frac{1}{N - m} \sum_{j=1}^{N-m} s_j^m \tag{33}$$

$$\mathbf{u}^{m+1} = \frac{1}{N - m} \sum_{j=1}^{N-m} s_j^{m+1}. \tag{34}$$

Finally, SE values are calculated using the expression:

$$SE = -\log\left(\frac{\mathbf{u}^{m+1}}{\mathbf{u}^m}\right). \tag{35}$$

The recommended embedded dimension  $m$  and tolerance level  $r$  values are  $m = 2, r = 0.2 \times \sigma$  where  $\sigma$  is the standard deviation of the time series [32]. Thus, these parameters are used in this work.

### 3.2.4. Dispersion Entropy (DE)

DE was introduced in [33] to overcome PE and Sample Entropy (SE) limitations. SE is slow in computation especially for long time series, whereas PE disregards information of the amplitude values mean and amplitude variations [34]. DE measure can be obtained as follows. Given the time series signal  $\{x(1), x(2), \dots, x(N)\}$ , with length of  $N$ , let  $\mathbf{x}$  be mapped to  $\mathbf{y} = \{y(1), y(2), \dots, y(N)\}$  using the Normal Cumulative Distribution Function (NCDF) which is defined as:

$$\mathbf{y} = F(\mathbf{x}|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mathbf{x}} e^{-\frac{(t-\mu)^2}{2\pi\sigma^2}} dt \tag{36}$$

NCDF values are the probabilities that a random variable from the normal distribution, with mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of the total signal  $\mathbf{x}$ , is less than  $x_j$  value from the time series  $\mathbf{x}$ . Next, each  $y(j); j = 1, \dots, N$  is assigned a class from 1 to  $c$  linearly as follows.

$$z_j^c = \text{round}(c \cdot y(j) + 0.5) \quad j = 1, \dots, N. \tag{37}$$

This provides  $N$  members of the classified time series. Here other linear or non-linear methods can also be employed. Embedding vectors  $\mathbf{z}_i^{m,c}$  with dimension  $m$  and time delay  $\tau$  are then created:

$$\mathbf{z}_i^{m,c} = \{z_i^c, z_{i+d}^c, \dots, z_{i+(m-1)d}^c\}; \quad i = 1, 2, \dots, N - (m - 1)\tau. \tag{38}$$

The latter is mapped to a dispersion pattern  $\pi_{v_0 v_1 \dots v_{m-1}}$ , among  $c^m$  possible dispersion patterns, in that  $v_0 = z_i^c, v_1 = z_{i+d}^c, \dots, v_{m-1} = z_{i+(m-1)d}^c$ .

The dispersion probability of occurrence for each pattern is then calculated as follows.

$$p(\pi_{v_0 v_1 \dots v_{m-1}}) = \frac{\sum_{i \leq N-(m-1)\tau} 1_{u:\text{type}(u)=\pi_{v_0 v_1 \dots v_{m-1}}}(\mathbf{z}_i^{m,c})}{N - (m - 1)\tau} \tag{39}$$

Finally, the DE value is obtained based on the Shannon entropy formula as follows.

$$DE = - \sum_1^{c^m} p(\pi_{v_0 v_1 \dots v_{m-1}}) \cdot \log(p(\pi_{v_0 v_1 \dots v_{m-1}})) \tag{40}$$

The calculated DE value provides the level of spreading in a time series, which may be informative on the features of the different discharge signals. In this paper, the standard values  $m = 2$  and  $c = 3$  are followed [33].

### 3.3. Classification Algorithms

Supervised classification process could be outlined as follows. Provided a labelled training data set, the classifier model learns a mapping space which is able to predict the true labels of an unseen testing data set. In this paper, two different classification algorithms, MCSVM and RF, are employed and discussed as follows.

#### 3.3.1. Support Vector Machine (SVM) and Multi-Class SVM (MCSVM)

SVM is used in this work as a binary classifier that groups separately data from two classes only in the feature space by means of a hyperplane, whose distance from the nearest point of each class determines the margin [35]. Ideal SVM performance requires wide margin and thus an optimum separation between the classes. The basic SVM separates two classes using a linear kernel function as illustrated in Figure 2. There exist other kernels, such as quadratic, polynomial and Radial Basis Function (RBF), to adapt the nature of data. Optimisation techniques are recommended to choose the best kernel function, along with its parameter, to employ for a high performance. Here, a grid search method [36] is employed as an optimisation technique. This involves multiple training and testing of the MCSVM with a grid of all possible kernel functions and their parameters.

The implementation theory of SVM can be summarised as follow: Let  $\mathbf{x}$  be the data input and  $\mathbf{y}$  the associated labels with length  $L$ . It is assumed that the data points belong to two classes “1” and “-1”. Each data point is non-linearly mapped to a feature space separated by a hyperplane with the basic geometric equation:

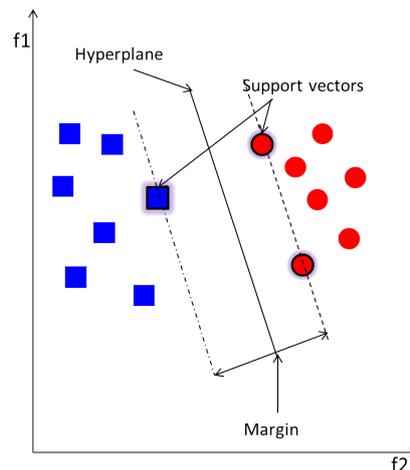
$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{41}$$

where  $b$  is a scalar and  $\mathbf{w}$  is L-dimensional vector parameters that play an important role to determine the hyperplane position. If  $b = 0$ , the hyperplane will pass by the origin. Otherwise, the margin is created or increased. The parallel hyperplanes that separate the two different data classes are defined as  $\mathbf{w} \cdot \mathbf{x} + b = 1$  and  $\mathbf{w} \cdot \mathbf{x} + b = -1$  for the first and second class respectively. The hyperplane is obtained as a solution to the optimisation problem in Equation (42), while considering the noise slack variable  $\zeta_i$  which determines the range to which the samples overstep the margin, and the error penalty  $c$  which represents the trade-off between maximisation of the margin and classification error on training phase.

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 + c \sum_{i=1}^M \zeta_i \tag{42} \\ \text{Subject to } & \begin{cases} \mathbf{y}_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 - \zeta_i \\ \zeta_i \geq 0, \end{cases} \quad i = 1, \dots, M \end{aligned}$$

where  $\zeta_i$  denotes the distance between the margin and the data point which is in error. The calculation of Equation (42) is simplified and solved by converting it to a Lagrangian problem which is explained in more detail in [37]. This introduces an  $\alpha_i$  parameter which expresses  $\mathbf{w}$  in solving Equation (42). The solution yields to a non-linear decision function expressed as:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i,j=1}^M \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_j) + b\right). \quad (43)$$



**Figure 2.** Support Vector Machine (SVM) linear separation.

The issues that may arise from a learning process in SVM with high dimension feature space are data over-fitting and computational errors. Over-fitting can be solved by introducing a kernel function  $\Phi(\mathbf{x})$  which performs a dot product of the feature space. The definition of this kernel function can be found in [35]. This is the case of non-linear mapping using a kernel function as discussed earlier for Figure 2. The non-linear vector function  $\Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_l(\mathbf{x}))$ , with feature space of dimension  $l$ , is implemented and the decision function in Equation (43) can be reformulated as:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i,j=1}^M \alpha_i y_i (\Phi^T(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) + b\right). \quad (44)$$

To classify more than two classes, the MCSVM approach is exploited by employing the one-against-one (OAO) strategy. This involves training each of  $k(k-1)/2$  models on two classes,  $p$  and  $q$ , as a normal binary classification provided that  $k$  is the total number of classes. A “Max-Win” voting method is employed for prediction. The vote for a predicted class is increased by one when the unseen data sample is close to that class, and the one that receives the highest vote is considered to be the predicted class of this data sample.

### 3.3.2. Random Forests (RF)

The RF classifier is a group of decision trees that are trained on random features sets of labelled data. The randomness property provides de-correlated trees. The model is formed by an ensemble of components including weak learners and leaf predictor type. The motivation for using RF revolves around its properties. The main one is the ability to handle more than two classes. Furthermore, the RF technique is efficient because of its low model variances (over-fitting) and its parallelism structure. These advantages are a motivation to employ RF as an additional classifier, then compare its performance against MCSVM.

In brief, training of an RF classifier is achieved in the following steps.

1. At an initial node, randomly choose  $\mathbf{P}$  feature instances from the overall instances  $\mathbf{Q}$  presented to the classifier, where  $\mathbf{P}$  is much smaller than  $\mathbf{Q}$ .
2. Calculate the best split point using Information Gain defined as:

$$I = H(\mathbf{s}) - \sum_{i \in \{1,2\}} \left| \frac{\mathbf{s}^i}{\mathbf{s}} \right| H(\mathbf{s}^i) \tag{45}$$

where  $H(\mathbf{s})$  is the Shannon Entropy [26] of the node  $\mathbf{s}$  and  $\mathbf{s}^i$  is the child node.

3. Using the best split point, divide the main node into daughter nodes and reduce the number of feature instances along the nodes.
4. Repeat steps 1 to 3 until a maximum depth  $l = 5$  is reached.
5. Repeat steps 1 to 4 for  $K = 500$  trees of the model. The more trees that are employed then the higher the achieved performance.

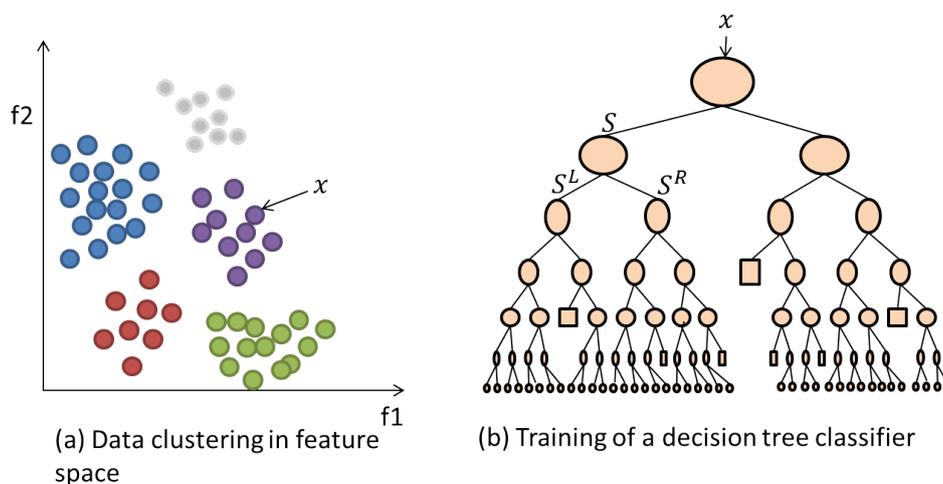
Let a training instance, from a training set  $\mathbf{Q} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ , be a vector of  $d$  dimension features  $\mathbf{x} = (x(1), \dots, x(d))$  with its associated labels  $\mathbf{y}$ . The overall RF model is an ensemble of  $K$  weak classifiers  $h = \{h_1(\mathbf{x}), \dots, h_K(\mathbf{x})\}$  where each  $h_k$  is a decision tree defined as:

$$h_k(\mathbf{x}) = h(\mathbf{x}|\gamma_k) \tag{46}$$

provided that  $\gamma_k = \gamma_{k1}, \dots, \gamma_{kp}$  is a set of parameters that decide the variables at which the nodes are split, tree structure etc. These parameters are optimised during the training phase by maximising the Information Gain objective function  $\gamma_j$  as:

$$\gamma_j^* = \operatorname{argmax}_{\gamma_j} \{I_j\} \tag{47}$$

where  $I_j$  is the Information Gain at node  $j$ . The mathematics of this algorithm are detailed further in [38]. Figure 3 illustrates the training of a single tree in the RF on a training data set as discussed in the previous steps. The data/label pair instances are utilised to optimise the parameters within each node in order to provide a trained model. The testing phase involves label prediction of the unseen data features based on the rules and parameters generated during the RF model creation. Each trained decision tree  $h_k$  within the model provides a prediction output and the highest voted label among all trees is considered to be the final predicted label.



**Figure 3.** (a) Example feature space representation of data instances that belong to 5 different classes (colours). (b) An example decision tree classifier architecture.

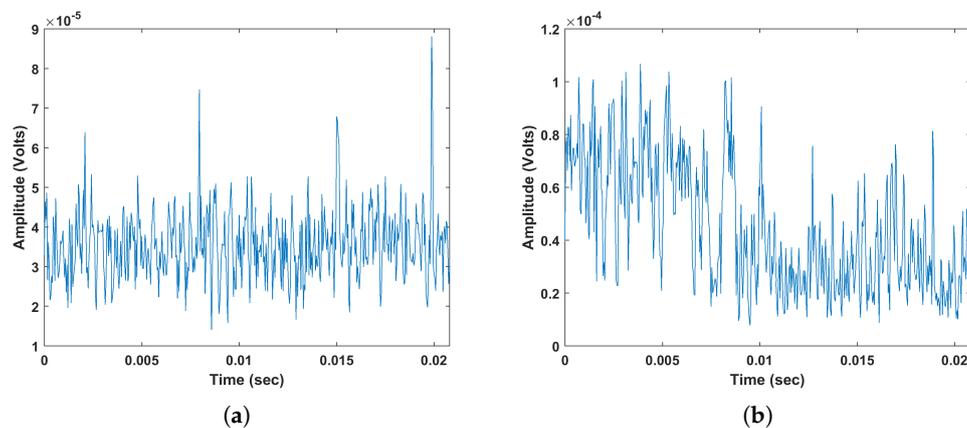
## 4. Experimental Set-Up

### 4.1. EMI Signals Measurement

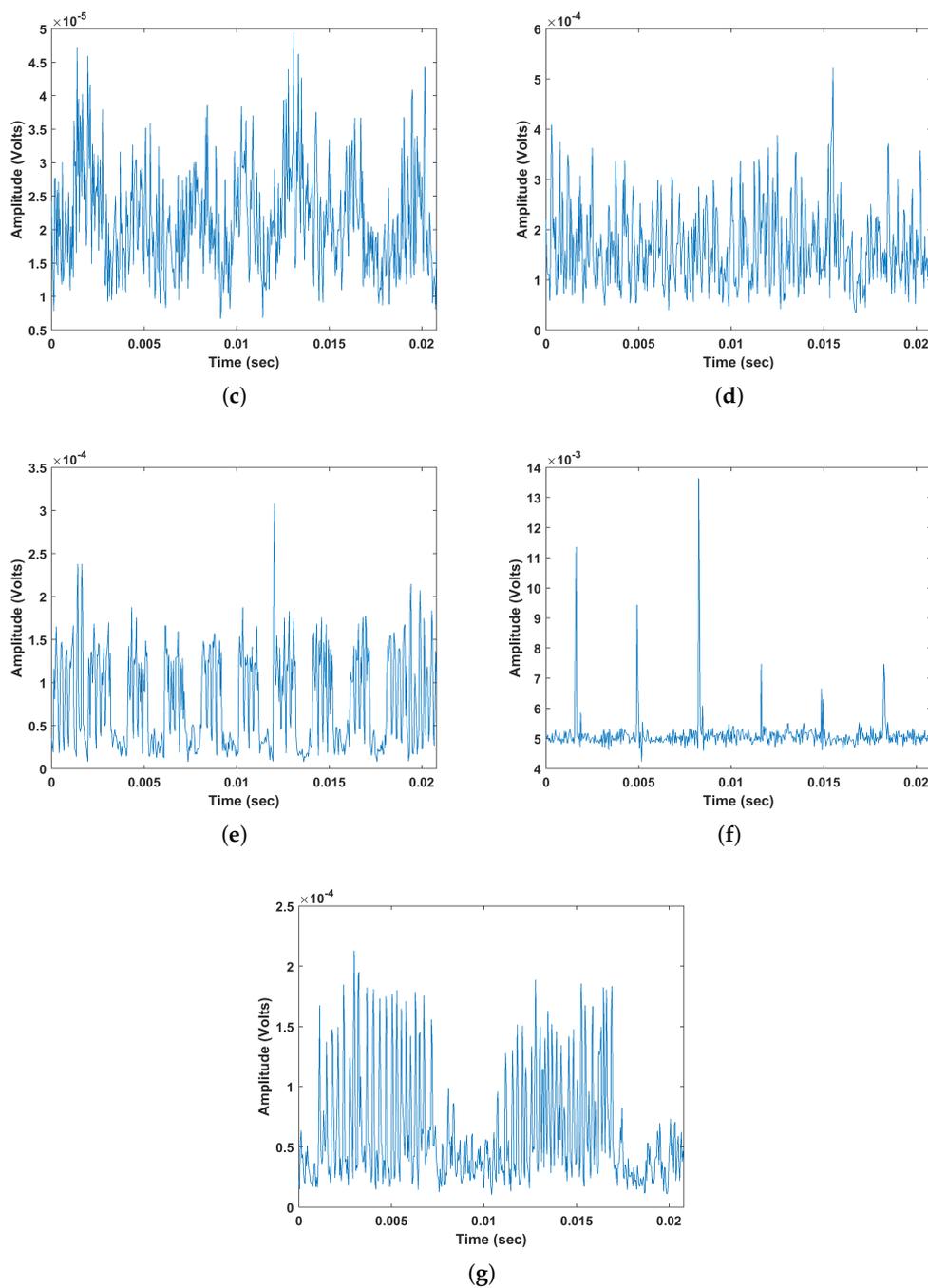
EMI measurement technique, described in Section 2, has been used for data acquisition and expert analysis. EMI data was collected from ten different operating HV sites at various assets including motors, generators, cables, transformers and isolated phase-bus. A total of 7 EMI discharge types were identified within the data known and abbreviated as PD, Arcing (A), Process Noise (PN), Random Noise (RN), Data Modulation (DM), Exciter (E), and micro Sparking (mS). Table 1 shows the discharge sources identified within each site, and Figure 4 shows example time series from each signal type. It is important to highlight that the primary contribution in this paper is about classification of the different discharges and not trending of their severity levels. Yet, when the fault is classified by the algorithm, trending could be performed, for instance, increase in repetition rate and magnitude level, can be assessed [17,39].

**Table 1.** Identified discharge sources per site.

Site	Discharge Source
1	PD, RN, PN
2	mS, DM, RN, PD, A
3	PD, E
4	PD, E
5	RN, DM, PD
6	RN, DM, E, PD, mS
7	PN, E, PD
8	PD, E
9	PN, E, PD
10	PD, E



**Figure 4.** Cont.



**Figure 4.** Time series signals of (a) Partial Discharge (PD) (b) Arcing (A) (c) Process Noise (PN) (d) Random Noise (RN) (e) Data Modulation (DM) (f) Exciter (E) (g) microSparking (mS).

#### 4.2. Application of Feature Extraction and Classification

The collected and labelled EMI signals were split into segments of 4000 samples for ease of feature computation. PE, WPE, DE and SE measures were applied to each segment which significantly extracts important characteristics of each EMI source while potentially reducing data dimension for the classification algorithms. A total of  $N \times 4$  instances were implemented in MCSVM and RF classifiers. Training of each classifier is performed by introducing the training data set, with its associated labels, to the classifier. This results in a trained model which can be tested by presenting unseen data set without labels. The label for each testing instance are predicted by the model, and then compared to

the true labels provided by “EMI experts”. The classifiers performance is obtained by calculating the accuracy (*acc*) %, precision (*pr*), recall (*r*) and F-measure (*F*) defined in Equations (48)–(51) respectively, where *tp* =true positives, *fp* = false positives, *fn* = false negatives and *tpr* = total predictions. Precision measures the level of exactness, whereas recall indicates the level of completeness within the classifier. F-measure is the harmonic mean of precision and recall which denotes the balance between them. A high value of these measures is preferable and the maximum performance has a value of 1. The precision of each class is presented in the confusion matrix (bottom row of matrices) and recall is shown in the last column accordingly.

$$acc = \frac{tp}{tpr} \times 100 \quad (48)$$

$$pr = \frac{tp}{tp + fp} \quad (49)$$

$$r = \frac{tp}{tp + fn} \quad (50)$$

$$F = 2 \times \frac{pr \times r}{pr + r} \quad (51)$$

In this paper ten fold cross-validation approach is utilised to evaluate performance consistency. This involves training the classifier ten times, each with a different 90% of the data instances. The remaining instances are used for testing the classifiers.

## 5. Results

The average classification accuracy from each fold within individual sites is calculated and presented in Table 2. The proposed approach demonstrated high performance using both MCSVM and RF classifiers, while both strongly compete with each other. It is observed that in 8 of the sites (sites 1 to 3 and 5 to 9) both classifiers perform similarly well with high accuracy, but one marginally outperforms the other, with the exception of site 7 using RF where the lowest accuracy is observed. Notably, in sites 4 and 10 both classifiers achieve the top classification accuracy of 100%.

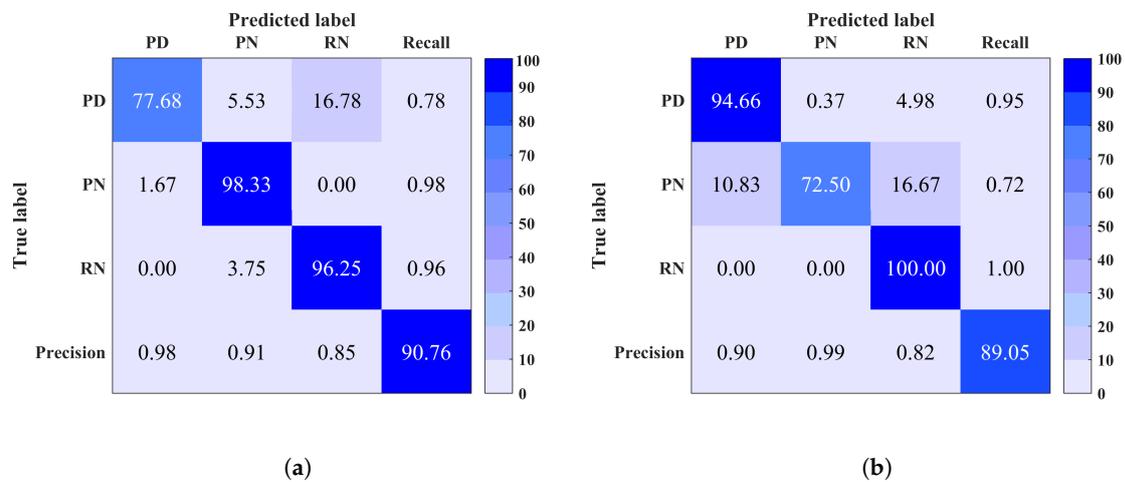
**Table 2.** Classification accuracy (rounded) results using Multi-Class Support Vector Machine (MCSVM) and Random Forest (RF).

Site		1	2	3	4	5	6	7	8	9	10
Accuracy %	MCSVM	91	75	91	100	96	99	100	99	100	100
	RF	89	79	92	100	97	98	72	100	99	100

To further investigate and understand the lower performance results i.e., accuracy < 98%, considering that a performance above 98% is excellent, the confusion matrix is calculated. It is important to visualise the classification algorithm’s performance from the confusion matrix aspect, especially for multi-class problems, as the total classification accuracy on its own may be misleading. The confusion matrix provides an in depth understanding on the predictions and errors. It shows the number of correct and wrong predictions within each class and what classes are being confused. Further discussion on these results is detailed as follows.

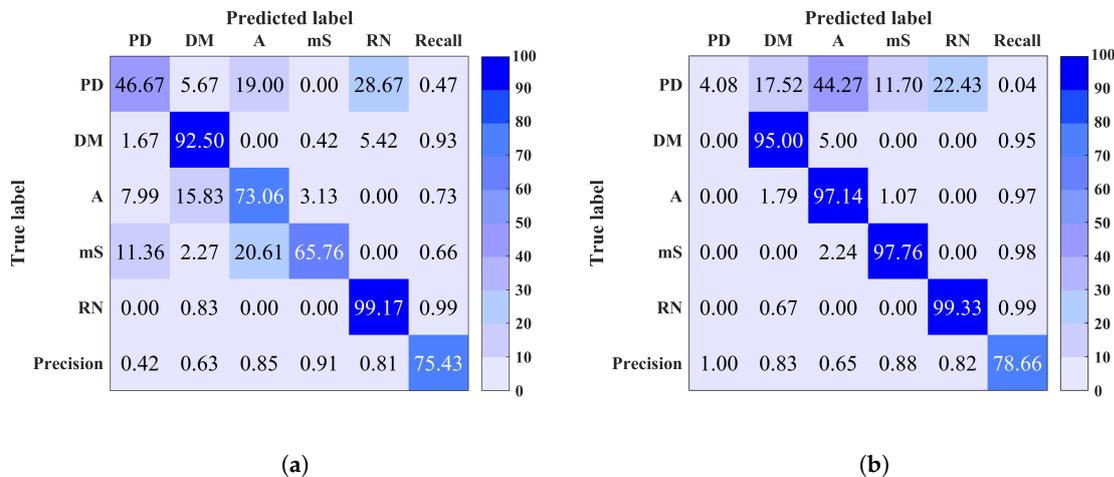
Site 1: Based on Figure 5, it is observed that although the total classification accuracy (shown in bottom right corner) is high, there is a confusion between PD, RN and PN. This could be due to the presence of high noise level within PD signals. In fact, the peaks of PD signals retrieved from high frequency in the EMI frequency spectrum tend to get attenuated and overwhelmed by noise. By referring to Figure 4 shows two example signals of PD and RN, in (a) and (b), collected from site

1. It is clear that PD signal is noisy which makes its classification more challenging, thus it could be easily confused with noise signals.



**Figure 5.** Confusion matrix of site 1 using (a) Multi-Class Support Vector Machine (MCSVM). (b) Random Forest (RF). Overall classification accuracy is shown in bottom right corner.

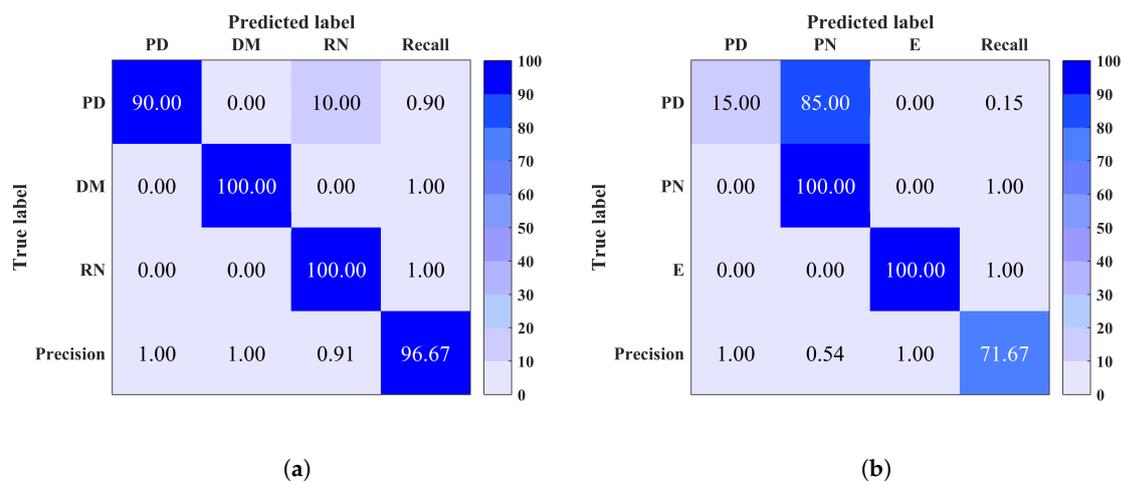
Site 2: The classification accuracy of this site is acceptable; however, it appears from Figure 6 that PD is mainly predicted as RN and A. It is likely that this confusion is due to the similar cause which was discussed for site 1.



**Figure 6.** Confusion matrix of site 2 using (a) Multi-Class Support Vector Machine (MCSVM). (b) Random Forest (RF). Overall classification accuracy is shown in bottom right corner.

Site 3: It is observed that this site contains two EMI classes only which are PD and E. Therefore the classification accuracy is affected by the confusion of these two classes only. This could be due to the similarity between PD and E signal nature i.e., narrow pulses shape.

Site 5: The obtained total classification accuracy for this site is very high using both classifiers. The confusion matrix in Figure 7a shows that the loss in classification accuracy is due to the prediction of some PD instances as RN. This is similar to the case of site 1 and 2.



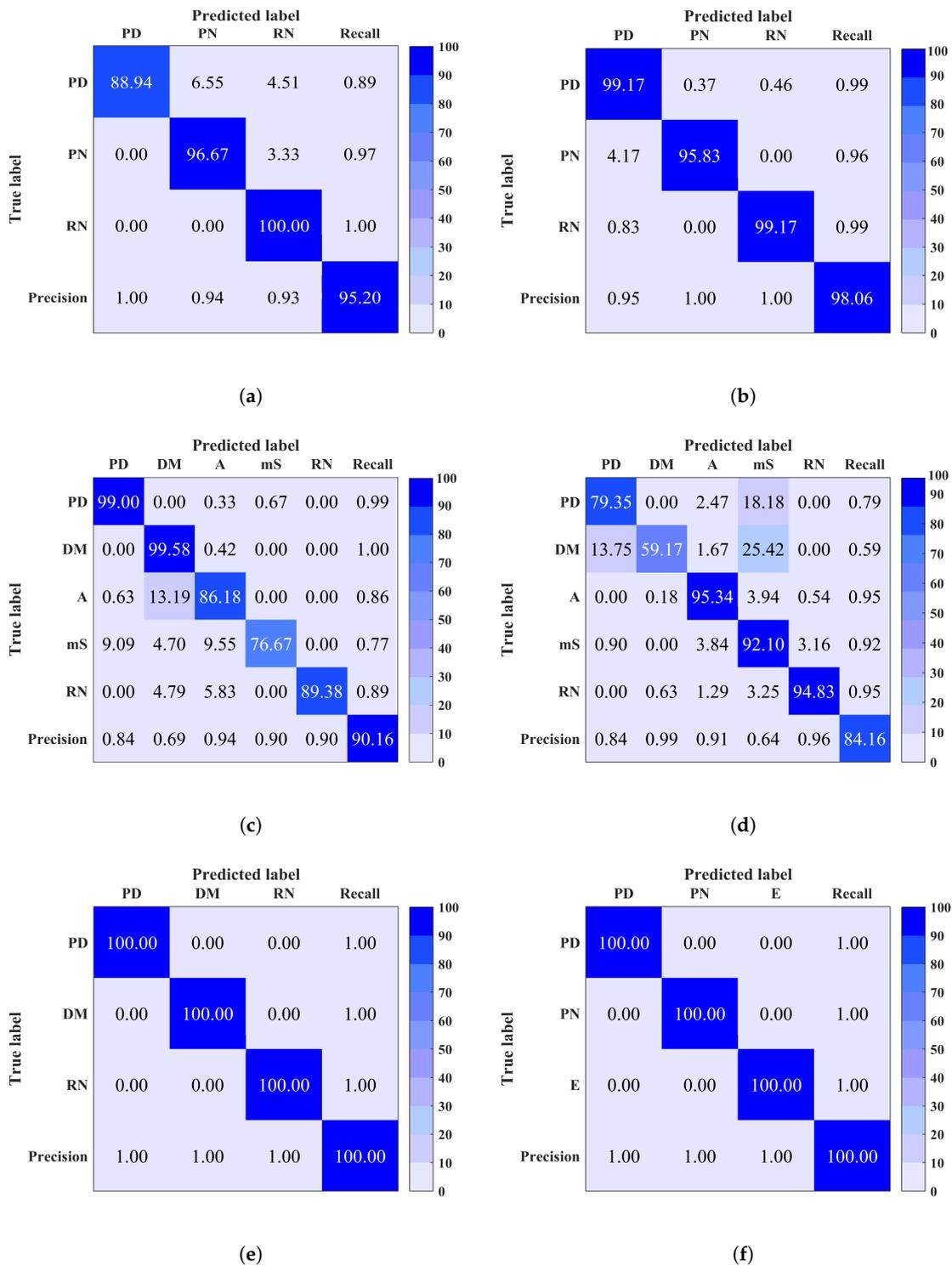
**Figure 7.** Confusion matrix of (a) site 5 (b) site 7 using Random Forest (RF). Overall classification accuracy is shown in bottom right corner.

Site 7: The confusion matrix of this site, illustrated in Figure 7b, explains the low classification accuracy such that PD instances were predicted as PN. Again, this is a similar case to sites 1, 2 and 5, where noisy PD signals are confused with noise. Note that for each site it is observed that the precision and recall are directly proportional to the classification rate within each class.

As shown above, noise could be a potential issue that may disrupt the classification performance. To overcome this issue, denoising using ALIF-TV technique is implemented as preprocessing.

Classification results after denoising of any input signal with PAPR < 15 dB are shown in Figure 8. A remarkable improvement in classification performance is achieved after applying denoising in each site (1, 2, 5 and 7). A total accuracy improvement of 4–9% was achieved in site 1 for both MCSVM and RF respectively. In site 2, an improvement of 5–15% for RF and MCSVM respectively, was obtained. The maximum accuracy of 100% was obtained in sites 5 and 7. Furthermore, an improvement in precision, recall and F-measure was also observed in each site. Table 3 summarises the average accuracy, precision, recall and F-measure for the sites to which denoising was applied, using both classifiers. The difference in performance between denoised and noisy signals classification is clear. In summary, denoising should be considered for a better prediction as it minimises the confusion between noise signals (RN and PN) and the noisy discharge sources signals.

A second case of classification was studied where the data from all ten sites was mixed. Classification results employing both MCSVM and RF classifiers, for the raw as well as denoised signals, are presented in Table 4. It is observed that MCSVM performs better in this case, particularly after denoising.



**Figure 8.** Confusion matrix of site 1 (a) using Multi-Class Support Vector Machine (MCSVM) (b) Random Forest (RF), site 2 (c) using MCSVM (d) using RF, and (e) site 5 (f) site 7 using RF, after denoising. Overall classification accuracy is shown in bottom right corner.

**Table 3.** Average classification performance measures before and after denoising using MCSVM/RF.

<b>Before Denoising</b>					
	<b>Site</b>	<b>1</b>	<b>2</b>	<b>5</b>	<b>7</b>
Accuracy %		91/89	75/79	96/97	100/72
Precision		0.91/0.90	0.72/0.84	0.96/0.97	1/0.85
Recall		0.91/0.89	0.76/0.79	0.97/0.96	1/0.72
F-measure		0.91/0.90	0.74/0.81	0.96/0.96	1/0.78
<b>After denoising</b>					
Accuracy %		95/98	90/84	100/100	100/100
Precision		0.96/0.98	0.85/0.87	1/1	1/1
Recall		0.98/0.98	0.90/0.84	1/1	1/1
F-measure		0.97/0.98	0.87/0.85	1/1	1/1

**Table 4.** Average classification performance measures for all sites before and after denoising using MCSVM/RF.

<b>Before Denoising</b>	<b>Accuracy %</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
	77/73	0.83/0.77	0.77/0.73	0.78/0.72
<b>After denoising</b>				
	91/66	0.91/0.79	0.91/0.66	0.91/0.65

## 6. Conclusions

This paper introduces the application of four entropy measures, called PE, WPE, DE and SE, as robust feature extraction techniques in the classification of 7 different EMI signal types by means of two popular classification algorithms known as RF and MCSVM. The developed model provides an intelligent system that captures an expert's knowledge on electrical machine condition assessment using the EMI measurement technique. The performance of both classifiers demonstrated an overall good classification accuracy. However, it was found that noise affects the prediction results. As a solution to this issue, a denoising approach was proposed for the noisy signals which are then selected based on their PAPR. Denoising is implemented prior to feature extraction and classification. This approach achieved an improvement in classification results. This work has successfully demonstrated the applicability of Entropy-based methods within EMI HV condition monitoring. The application of other Entropy-based methods, such as Tsallis and Renyi will be considered in future work. In summary, the proposed feature extraction method transfers the measured complex time series signals to a simple low dimension data set which facilitates the classification process. This work could be considered in future developments of an automatic condition monitoring instrument for industrial applications, based on the fact that entropy measures are easy and simple to compute.

**Author Contributions:** All authors on this paper contributed to the work. I.M., G.M., B.G.S. and A.N. conceived and designed the experiments and helped draw conclusions; I.M. and performed the experiments; I.M., G.M. and A.N. analyzed the data; P.B. contributed reagents/materials/analysis tools; I.M. wrote the paper and G.M. and B.G.S. provided corrections.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kuffel, E.; Zaengl, W.S.; Kuffel, J. *High Voltage Engineering Fundamentals*, 2nd ed.; Newnes: Oxford, UK, 2000; pp. 8–76.

2. Kreuger, F.H. *Industrial High Voltage: 4. Coordinating, 5. Testing, 6. Measuring*; Delft University Press: Delft, The Netherlands, 1992; pp. 117–129.
3. Bodega, R.; Morshuis, P.H.F.; Lazzaroni, M.; Wester, F.J. PD Recurrence in Cavities at Different Energizing Methods. *IEEE Trans. Instrum. Meas.* **2004**, *53*, 251–258. [[CrossRef](#)]
4. Robles, G.; Parrado-Hernandez, E.; Ardila-Rey, J.; Martinez-Tarifa, J.M. Multiple partial discharge source discrimination with multiclass support vector machines. *Expert Syst. Appl.* **2016**, *55*, 417–428. [[CrossRef](#)]
5. Spyker, R.; Schweickart, D.L.; Horwath, J.C.; Walko, L.C.; Grosjean, D. An evaluation of diagnostic techniques relevant to arcing fault current interrupters for direct current power systems in future aircraft. In Proceedings of the Electrical Insulation Conference and Electrical Manufacturing Expo (EICEME'05), Indianapolis, IN, USA, 23–26 October 2005; pp. 146–150.
6. Tang, W.H.; Wu, Q.H. *Condition Monitoring and Assessment of Power Transformers Using Computational Intelligence*; Springer-Verlag: Liverpool, UK, 1992; pp. 1–13.
7. Sureshjani, S.A.; Kayal, M. A Novel Technique for Online Partial Discharge Pattern Recognition in Large Electrical Motors. In Proceedings of the IEEE 23rd International Symposium on Industrial Electronics (SIE'14), Istanbul, Turkey, 1–4 June 2014; pp. 721–726.
8. Hunter, J.A.; Lewin, P.L.; Hao, L.; Walton, C.; Michel, M. Autonomous classification of PD sources within three-phase 11 kV PILC cables. *IEEE Trans. Dielectr. Electr. Insul.* **2013**, *20*, 2117–2124. [[CrossRef](#)]
9. Majidi, M.; Fadali, M.S.; Etezadi-Amoli, M.; Oskuoee, M. Partial discharge pattern recognition via sparse representation and ANN. *IEEE Trans. Dielectr. Electr. Insul.* **2015**, *22*, 1061–1070. [[CrossRef](#)]
10. Zhang, S.; Li, C.; Wang, K.; Li, J.; Liao, R.; Zhou, T.; Zhang, Y. Improving recognition accuracy of partial discharge patterns by image-oriented feature extraction and selection technique. *IEEE Trans. Dielectr. Electr. Insul.* **2016**, *23*, 1076–1087. [[CrossRef](#)]
11. Karthikeyan, B.; Gopal, S.; Vimala, M. Conception of complex probabilistic neural network system for classification of partial discharge patterns using multifarious inputs. *Expert Syst. Appl.* **2005**, *29*, 953–963. [[CrossRef](#)]
12. Hunter, J.A.; Hao, L.; Lewin, P.L.; Evagorou, D.; Kyprianou, A.; Georghiou, G.E. Comparison of two partial discharge classification methods. In Proceedings of the IEEE International Symposium on Electrical Insulation Conference (ISEI'10), San Diego, CA, USA, 6–9 June 2010; pp. 1–5.
13. Timperley, J.E. Comparison of PDA and EMI diagnostic measurements [for machine insulation]. In Proceedings of the Conference Record of the 2002 IEEE International Symposium on Electrical Insulation, Boston, MA, USA, 7–10 April 2002; pp. 575–578.
14. Timperley, J.E.; Vallejo, J.M. Condition Assessment of Electrical Apparatus With EMI Diagnostics. *IEEE Trans. Ind. Appl.* **2017**, *53*, 693–699. [[CrossRef](#)]
15. Timperley, J.E. Audio spectrum analysis of EMI patterns. In Proceedings of the 2007 Electrical Insulation Conference and Electrical Manufacturing Expo, Nashville, TN, USA, 22–24 October 2007; pp. 39–41.
16. International Special Committee on Radio Interference. *IEC CISPR 1-6-1-1:2015*; IEC: Geneva, Switzerland, 2015.
17. Timperley, J.E.; Vallejo, J.M.; Nesbitt, A. Trending of EMI data over years and overnight. In Proceedings of the 2014 IEEE Electrical Insulation Conference (EIC), Philadelphia, PA, USA, 8–11 June 2014; pp. 176–179.
18. Feher, K. *Telecommunications Measurements, Analysis, and Instrumentation*; SciTech Publishing: Stevenage, UK, 1997; p. 188.
19. Mitiche, I.; Morison, G.; Nesbitt, A.; Hughes-Narborough, M.; Boreham, P.; Stewart, B.J. An Evaluation of Total Variation Signal Denoising Methods for Partial Discharge Signals. In Proceedings of the 13th International Electrical Insulation Conference (INSUCON), Birmingham, UK, 16–18 May 2017; pp. 1–5.
20. Ding, Y.; Selesnick, I.W. Artifact-Free Wavelet Denoising: Non-convex Sparse Regularization, Convex Optimization. *IEEE Signal Process. Lett.* **2015**, *22*, 1364–1368. [[CrossRef](#)]
21. Donoho, D.L. De-noising by soft thresholding. *IEEE Trans. Inf. Theory* **1995**, *41*, 613–627. [[CrossRef](#)]
22. Afonso, M.; Bioucas-Dias, J.; Figueiredo, M.T. Fast Image Recovering Using Variable Splitting and Constrained Optimization. *IEEE Trans. Image Process.* **2010**, *19*, 2345–2356. [[CrossRef](#)] [[PubMed](#)]
23. Eckstein, J.; Bertsekas, D. On the Douglas Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators. *Math. Program.* **1992**, *5*, 293–318. [[CrossRef](#)]
24. Bandt, C.; Pompe, B. Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys. Rev. Lett.* **2002**, *88*, 174102. [[CrossRef](#)] [[PubMed](#)]

25. Amigo, J.M.; Keller, K. Permutation Entropy: One Concept, two Approaches. *Eur. Phys. J. Spec. Top.* **2013**, *222*, 263–273. [[CrossRef](#)]
26. Rathie, P.; Da Silva, S. Shannon, Levy, and Tsallis: A note. *Appl. Math. Sci.* **2008**, *2*, 1359–1363.
27. Riedl, M.; Müller, A.; Wessel N. Practical Considerations of Permutation Entropy. *Eur. Phys. J. Spec. Top.* **2013**, *222*, 249–262. [[CrossRef](#)]
28. Yan, R.; Liu, Y.; Gao, R.X. Permutation Entropy: A nonlinear Statistical Measure for Status Characterization of Rotary Machines. *Mech. Syst. Signal Process.* **2012**, *29*, 474–484. [[CrossRef](#)]
29. Fadlallah, B.; Chen, B.; Keil, A.; Príncipe, J. Weighted-Permutation Entropy: A Complexity Measure for Time Series Incorporating Amplitude Information. *Phys. Rev. E* **2013**, *87*, 022911. [[CrossRef](#)] [[PubMed](#)]
30. Xia, J.; Shang, P.; Wang, J.; Shi, W. Permutation and Weighted-Permutation Entropy Analysis for the Complexity of Nonlinear Time Series. *Commun. Nonlinear Sci. Numer. Simul.* **2016**, *31*, 60–68. [[CrossRef](#)]
31. Richman, J.S.; Lake, D.E.; Moorman, J.R. Sample Entropy. In *Numerical Computer Methods, Part E*; Academic Press: San Diego, CA, USA, 2004; pp. 172–184.
32. Ahmed, M.U.; Mandic, D.P. Multivariate Multiscale Entropy: A tool for Complexity Analysis of Multichannel Data. *Phys. Rev. E* **2011**, *84*, 061918. [[CrossRef](#)] [[PubMed](#)]
33. Rostaghi, M.; Azami, H. Dispersion Entropy: A Measure for Time-Series Analysis. *IEEE Signal Proc. Lett.* **2016**, *23*, 610–614. [[CrossRef](#)]
34. Massimiliano, Z.; Luciano, Z.; Osvaldo, R. A.; Papo, D. Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review. *Entropy* **2012**, *14*, 1553–1577. [[CrossRef](#)]
35. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1961.
36. Boardman, M.; Trappenberg, T. A Heuristic for Free Parameter Optimization with Support Vector Machines. In Proceedings of the International Joint Conference on Neural Networks, Vancouver, BC, Canada, 16–21 July 2006; pp. 1337–1344.
37. Widodo, A.; Yang, B.-S. Support Vector Machine in Machine Condition Monitoring and Fault Diagnosis. *Mech. Syst. Signal Process.* **2007**, *21*, 2560–2574. [[CrossRef](#)]
38. Criminisi, A.; Konukoglu, E.; Shotton, J. *Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning*; Microsoft Technical Report; Microsoft Research: Washington, DC, USA, 2011; pp. 5–19.
39. Timperley, J.E. Generator condition assessment through EMI diagnostics. In Proceedings of the ASME 2008 Power Conference, Lake Buena Vista, FL, USA, 22–24 July 2008; pp. 349–354.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).