sustainability

MDPI

Article

Extended Isolation Forests for Fault Detection in Small Hydroelectric Plants

Rodrigo Barbosa de Santis [1,*] and Marcelo Azevedo Costa [1,2]

[1] Graduate Program in Industrial Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte 31270-901, MG, Brazil; azevedo@est.ufmg.br

[2] Department of Industrial Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte 31270-901, MG, Brazil

* Correspondence: rsantis@ufmg.br; Tel.: +55-31-3409-4881

Received: 29 June 2020; Accepted: 29 July 2020; Published: 10 August 2020

Abstract: Maintenance in small hydroelectric plants is fundamental for guaranteeing the expansion of clean energy sources and supplying the energy estimated to be necessary for the coming years. Most fault diagnosis models for hydroelectric generating units, proposed so far, are based on the distance between the normal operating profile and newly observed values. The extended isolation forest model is a model, based on binary trees, that has been gaining prominence in anomaly detection applications. However, no study so far has reported the application of the algorithm in the context of hydroelectric power generation. We compared this model with the PCA and KICA-PCA models, using one-year operating data in a small hydroelectric plant with time-series anomaly detection metrics. The algorithm showed satisfactory results with less variance than the others; therefore, it is a suitable candidate for online fault detection applications in the sector.

Keywords: hydroelectric power plant; condition-based maintenance; machine learning; early fault detection; decision tree algorithm

1. Introduction

With energy demand expected to double by 2060, the development of clean energy sources is essential for guaranteeing an energy supply in the coming decades. Renewable energy already represents three quarters of yearly new installed capacity [1], and those related to water resources are the most applied. In this group, the construction of small hydroelectric plants (SHPs) has grown worldwide due to the lower initial investment, low operating costs, and increasing regulation of energy markets. The potential total energy generation capacity of these SHPs is twice the current total capacity of the presently installed energy plants [2].

Several case studies are reported in recent literature addressing the energy potential and importance of developing SHPs in emerging countries like Brazil [3], Turkey [4], Nigeria [5], and other sub-Saharan African [6] countries. Overall life cycle assessment is applied for quantitative economic evaluation of this type of undertaking in India [7] and Thailand [8]. Economic models of viability sensitivity analysis of SHPs stations are presented and applied to the energy context in Spain [9] and Greece [10]. A common factor among all these models of economic viability is the cost of operation and maintenance, which is a determining variable for the development of new stations.

However, maintenance of a hydroelectric generating plant is a complex task. It requires a certain level of expertise to ensure a satisfactory level of reliability of the asset during its useful life. There are three types of maintenance. The first, and most rudimentary, is corrective maintenance, in which a component is expected to break, and is then replaced. Preventive maintenance estimates the expected service life of a component, and replacement is done when the operating lifetime is reached.

Finally, in predictive maintenance, the condition of the system is calculated from data periodically or continuously obtained from various sensors [11,12]. A predictive, or condition-based maintenance (CBM) system, consists of two main phases. The first phase is diagnosis, which comprises fault detection or abnormal operating conditions, fault isolation by sub-components, and identification of the nature and extent of the failure [12]. The next phase is prognosis, applying statistical and machine learning models to estimate the useful life of the equipment and the confidence interval of the prediction [13], to anticipate maintenance, and to increase the reliability and availability of the generation system.

Examples of commonly applied methods for estimating useful life are divided between statistics [14], which includes the regression methods, Wiener process, gamma process, based on Markovian processes; machine learning methods such as neural networks, vector support machine, electrical signature analysis [15], principal component analysis [16]; and, more recently, deep learning techniques such as auto-encoder, recurrent, and convolution neural networks [17]. Several other energy sectors already make extensive use of these techniques: nuclear [18–20], wind [21,22], and solar [23,24].

Multivariate statistical methods such as Principal Component Analysis (PCA) [25], Independent Component Analysis (ICA) [26] and Least Square Support Vector Machine (LS-SVM) [27–30], have been widely applied for fault detection and diagnosis in hydro-generating systems. For instance, PCA decomposition is applied to aid experts in identifying and selecting the main features which contribute to cavitation in hydro-turbines [31]. Recent studies have proposed a new monitoring method, based on ICA-PCA that can extract both non-Gaussian and Gaussian information of process data for fault detection and diagnosis [32]. Later, the ICA-PCA was extended with the adoption of a nonlinear kernel transformation prior to the application of the decomposition method, which became known as the Kernel ICA-PCA (KICA-PCA) [33]. They reported its application in the hydroelectric generation context with higher success rates and lower fault detection delays than either the PCA or ICA-PCA applications.

The isolation forest (iForest) [34,35] is an anomaly detection model based on decision trees which, recently, is appearing in several case studies of anomaly detection in the business [36], industrial [37], and virtual security [38,39] areas. Briefly, the iForest method provides a non-parametric density estimate of the data. The non-parametric density can be estimated using data under normal operating conditions. After fitting the iForest model, density estimates or anomaly scores are calculated using online data. Faults are detected when the anomaly scores are higher than a pre-defined upper bound, indicating that the system under monitoring is no longer operating under normal conditions. The meta-heuristic model has some interesting advantages when compared to the classical linear decomposition models: it can handle an enormous amount of data and heterogeneous variables, without needing a data labeling process. It can, thus, develop nonlinear models of learning based on random, decision tree ensembles.

The most recent version of the algorithm, the extended isolation Forest (EIF), adopts hyperplanes with random slopes to separate the data, solving problems related to how the algorithm calculates the anomaly score [40]. The EIF can build scores with less variance and obtain better accuracy in the area under the receiver operating characteristics metric, compared to the original algorithm, without sacrificing computational efficiency [36,40]. However, no study has been found reporting the application of the iForest or EIF in hydroelectric turbines.

In this context, the present paper proposes the application of iForest and EIF to support fault detection and diagnosis of a hydro-generating unit (HGU) in an SHP. We compared the algorithms with PCA and KICA-PCA, using specific metrics for anomaly detection in time series [41]. The main findings and contributions of the current paper are:

- Application of iForest and EIF for intelligent fault diagnosis in an SHP generating unit.
- Proposal of the application of time distance and count detection metrics, most appropriate for the evaluation of models in the context of anomalies detection in time series.

- EIF presented reductions of 40.62% and 7.28% in the temporal distance, compared to the PCA and KICA-PCA.

The remainder of the present article is organized as follows. Section 2 defines the study methodology, describing the methods, algorithms, and data set applied. Section 3 presents the results and discussions of the simulations of the models, in addition to the outputs of the forest committee with illustrative examples of imminent failures. Finally, Section 4 presents the conclusions and recommendations for future work.

2. Materials and Methods

2.1. Dataset

The current study was developed in Ado Popinhak, an SHP situated in the southern region of Brazil. With an installed capacity of 22.6 MW, the plant supplies energy to 50,000 residences. Condition monitoring data from the main single HGU are registered every 5 min, and the scope of the study period is from 13 August 2018 to 9 August 2019.

We filtered out from the dataset the periods of maintenance, planned stop, operator intervention, or another status not associated with normal operation. Event data related to the asset are gathered from status reports and the maintenance management system to identify past failures. Fifty-nine faults were registered in the disclosed period, totaling 123 h of downtime. Six monitored variables are used: generator apparent power; bearing hydraulic lubrication unit (HLU) inflow; and, bearing vibration from four different positions: axial, vertical radial, horizontal radial, and coupled.

Figure 1 presents the interaction between the variables in the dataset, in two different visualizations. The vibration variables are replaced by the average of the variables at the different measuring points. Figure 1a indicates a low-apparent power region, where the average vibration is higher than in the rest of the observations. These present a transient period in which the generator unit operates with imbalanced water inflow inside the runner. In such a state, the wear damage to the system and the fault risks are more serious.

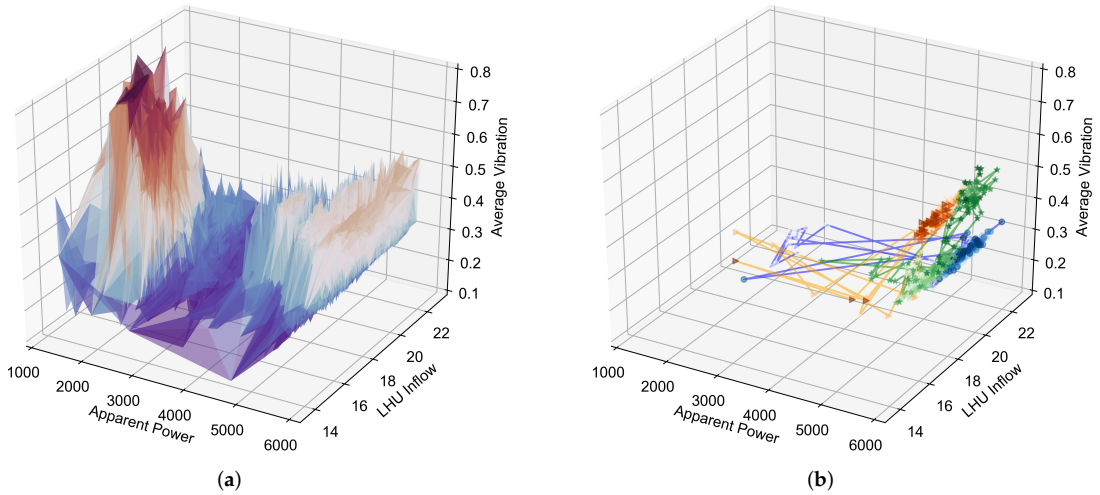


Figure 1. Graphical representation of the data set in three dimensions. In (**a**), the entire data set is presented, regardless of the temporal relationship between data points. In (**b**), three excerpts from the series with imminent failures are presented, each in a different color. The darker the marker, the closer the fault. Points connected by lines represent sequential states.

Figure 1b presents three excerpts of the time series before failure. The analysis of the representation indicates that the failures generally occur in regions where the vibration and apparent power are at

their maximum, and there may be significant fluctuations in the power and flow of HLU before they occur. The figure presents only a sub-sample, of 3 out of the 59 faults in the entire database, to avoid overload of information in the representation, which would make it difficult for the reader to analyze.

A fixed period, from 12 h prior to the failure up to the failure, splits the full data set into a training set and a test set. The training set corresponds to the healthy state, or normal operation, as long as the test set is linked to abnormal operation. In this way, the algorithm focuses its training on the positive class related to normal operating conditions, thus becoming a density estimator of the class of interest [42]. This type of approach is common in problems of unbalanced classes, in the context of anomaly detection, in which negative cases (our outliers) are absent or not adequately sampled [43].

After separation, the training and test set sizes were 47,857 and 4897, respectively. The ratio between training and test sets is about 10:1, which is an appropriate ratio when compared to reference studies on failure detection in hydroelectric plants that have adopted proportions of 8:1 [44] and 1140:100∼10:1 [33]. Anomaly detection algorithms, reported sequentially, are trained using only training data.

2.2. PCA

Principal Component Analysis (PCA) is a linear decomposition technique, effective for data dimensionality reduction that projects the correlated variables onto smaller sets of new variables that are orthogonal and retain most of the original variance. PCA is the most widely used data-driven technique for process monitoring, due to its capacity to deal with high-dimensional, noisy, and correlated data variance [45].

Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be an observation matrix, where n is the number of samples, and m is the number of monitor variables. $\mathbf{X}$ can be decomposed by the function

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

where $\mathbf{E}$ is the residual matrix, $\mathbf{T} \in \mathbb{R}^{nxa}$ is the score matrix, and $\mathbf{P} \in \mathbb{R}^{mxa}$ is loading matrix. The measure of PCA variance can be obtained by Hotelling's T^2 statistic representing the sum of the normalized squared scores

$$T^2 = \mathbf{t}^T \mathbf{D}^{-1} \mathbf{t}, \tag{2}$$

where $\mathbf{D}$ is the diagonal matrix of the eigenvalues with the retained principal components and $\mathbf{t} = \mathbf{P}^T x$, is the score of PCA, calculated from the multiplication of each element x and the loading matrix $\mathbf{P}$.

The T^2 index is used for monitoring processing, detecting a systematic variation of the process every time an observation exceeds the confidence limit T_α^2, given by

$$T_\alpha^2 = \frac{(n^2-1)\alpha}{n(n-\alpha)} F_\alpha(\alpha, n-\alpha) \tag{3}$$

where n is the number of samples, α is the number of sensed variables, F_α is the upper 100% critical point of F-distribution with α and $n-\alpha$ degree of freedom. As for classification, a set of class labels C is set as 1 if $T_i^2 > T_\alpha^2$ or else 0, if conditions are not met, for $T_1^2, T_2^2, ..., T_n^2$.

2.3. KICA-PCA

The KICA-PCA method provides a kernel transformation of data into higher dimensional data, prior to the application of decomposition. Thus, the method is capable of handling nonlinear multivariate processes, such as SHP condition monitoring [33].

In this application, we adopted the explicit mapping to a low-dimensional Euclidean inner product space using a randomized feature map $z : \mathbb{R}^{nxm} \rightarrow \mathbb{R}^{nxd}$ proposed by [46], so that the inner product between a pair of transformed points approximates their kernel evaluation:

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle \approx z(x)'z(y). \tag{4}$$

Contrary to kernel's lifting Φ, z is low dimensional and k is the radial basis function $k(x, y) = exp(-||x - y||^2/\sigma)$ and σ is the standard deviation.

The z mapping competes favorably in speed and accuracy, as evidenced by [46–48], being capable of handling the large training matrix of this study without exceeding computational resources of a standard personal computer.

The transformed matrix $\mathbf{X}'$ is calculated by the kernel approximation $\mathbf{z}^T\mathbf{z}$, such as each element

$$k(x_i, x_j) = x'_{ij} = z(x_i)^T z(x_j), \tag{5}$$

where x_i and x_j are the ith and jth columns of $\mathbf{X}$, respectively.

Before the application of ICA, the matrix $\mathbf{X}'$ should be whitened to eliminate the cross-relations among random variables. One popular method for whitening is to use the eigenvalue decomposition, considering $x'(k)$ with its co-variance $R'_x = E\{x'(k)x'(k)^T\}$, as described in [32]. The association of the kernel transform and the ICA is known in the literature as KICA.

ICA is a statistical, computational technique originally proposed to solve the blind source separation problem by revealing patterns hidden in signals, variable sets, or measurements [32]:

$$\bar{\mathbf{X}}' = \mathbf{AS} + \mathbf{E} \tag{6}$$

where $\bar{\mathbf{X}}'$ is the the whitened transformed matrix, $\mathbf{A}$ is the mixing matrix, $\mathbf{S}$ is the independent component matrix, and $\mathbf{E}$ is the residual matrix. The basic problem is estimating the original component $\mathbf{S}$ and the matrix $\mathbf{A}$ from $\bar{\mathbf{X}}'$. ICA calculates a separating matrix $\mathbf{W}$ such that the components of the reconstructed matrix $\mathbf{S}$ become as independent of each other as possible, given as

$$\hat{\mathbf{S}} = \mathbf{W}\bar{\mathbf{X}}'. \tag{7}$$

From the multiplication of $x'' = \mathbf{S}^T\hat{x}'$, a new matrix $\mathbf{X}''$ is obtained which represents the independent components (ICs) from the sensed data. These matrices are used as input for the PCA monitoring algorithm in Equation (1) and used to calculate the T^2 score and classification set from the comparison with the T^2_α threshold.

2.4. iForest and EIF

While most anomaly detection approaches are based on normal instance profiling, iForest is an anomaly detection algorithm that explicitly isolates anomalies. The method exploits two particularities of anomalies: they represent fewer instances in the observed set, and, compared to healthy instances, they have discrepant attribute-values [35].

The method does not apply any distance or density measures, thereby eliminating the major computational cost of distance calculation. In addition, the algorithm scales up linearly while keeping memory usage low and constant, which aligns with parallel computing, making the model suitable for handling large, high-dimensional data sets.

The anomaly detection procedure using iForest is a two-stage procedure: the training stage constructs the isolation trees (iTree), using sub-samples from the training set; the subsequent evaluation stage calculates the anomaly score for each instance of test set [34,35].

The iForest builds an ensemble of binary trees individually trained using a sub-sample $\mathbf{X}^s$ randomly drawn from $\mathbf{X}$, $\mathbf{X}^s \subset \mathbf{X}$. There are two control parameters in Algorithm 1: (1) the

sub-sampling rate ψ sets the number of samples used for each tree training, and (2) the number of trees nt, related to the complexity of the model.

Algorithm 1: *iForest* (**X**, nt, ψ)

Input: **X**—input data, nt—number of trees, ψ—sub-sampling size.
Output: a set of t *iTrees*
Initialize *Forest*

for $i \leftarrow 1$ **to** nt **do**
  $\mathbf{X}^s \leftarrow sample(\mathbf{X}, \psi)$;
  $Forest \leftarrow Forest \cup iTree(\mathbf{X}^s)$;
end
return Forest

The normal points tend to be isolated at the deeper end of the tree, whereas anomalies are closer to the tree root, due to their singularity nature. The shorter the average path length, the higher the chances to be anomalies. Hence, the anomaly score s is then defined by:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{8}$$

where n is the number of samples in the dataset, $E(h(x))$ is the average of path length $h(x)$ from a group of isolation tree, and $c(n)$ is the average of $h(x)$ given n, used for normalizing the path length. If an instance returns an anomaly score s very close to 1, it is very likely that one represents an anomaly; if it is much smaller than 0.5, it is safe to say the instance is normal; if the instance returns $s \approx 0.5$, the sample does not present any distinct anomaly [34].

Although the standard iForest algorithm is computationally efficient, there is a limitation as to how the anomaly score aggregates tree branches' length. Branch cuts are always horizontal or vertical, which introduces a bias in the anomaly score map.

The EIF algorithm can overcome this limitation by adopting random slopes along with the branching process. The selection of the branch cut then requires a random slope and a random intercept chosen from the range of values available in the training data. Each random slope is drawn from a random number for each coordinate of a vector $\vec{m}$ of size equal to the number of variables in a normal distribution $N(0, 1)$. The intercept is obtained from the uniform distribution of a range of values present at the branching point. The splitting criterion for a point x is given by $(\vec{x} - \vec{p}).\vec{m} \leq 0$.

Algorithm 2: *iTree* (**X**, e, hl)

Input: **X**—input data, e—current tree height, hl—height limit
Output: an iTree
if $e \geq hl$ or $|\mathbf{X}| \leq 1$ **then**
  return $exNode\{Size \leftarrow |\mathbf{X}|\}$
else
  randomly select a normal vector $\vec{m} \in \mathbb{R}^{|\mathbf{X}|}$ by drawing each coordinate of $\vec{m}$ from a standard Gaussian distribution.
  randomly selects an intercept point $\vec{p} \in \mathbb{R}^{|\mathbf{X}|}$ in the range of **X**
  set coordinates of $\vec{m}$ to zero according to extension level
  $\mathbf{X}_{hl} \leftarrow filter(\mathbf{X}, (\mathbf{X} - \vec{p}).\vec{m} \leq 0)$
  $\mathbf{X}_r \leftarrow filter(\mathbf{X}, (\mathbf{X} - \vec{p}).\vec{m} > 0)$
  return $inNode\{Left \leftarrow iTree(\mathbf{X}_{hl}, e + 1, el), Right \leftarrow iTree(\mathbf{X}_r, e + 1, el),$
    $Normal \leftarrow \vec{m}, Intercept \leftarrow \vec{p}\}$
end

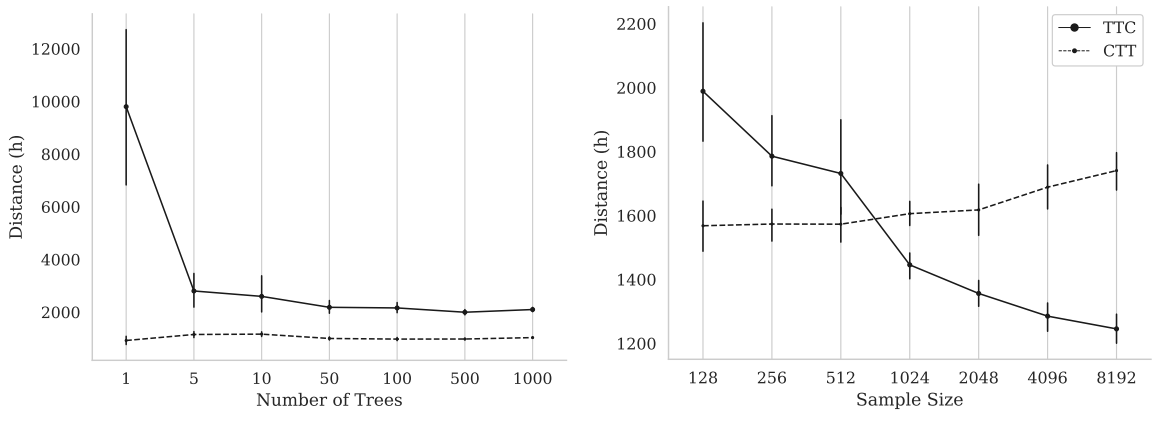
The property of concentration of data in clusters is maintained with the algorithm, as the intercept points $\vec{p}$ tend to accumulate where the data are, while the score maps are free of previously observed artifacts. EIF implementation modifies the lines of the original formulation in Algorithm 2 that describes the choice of the random value and intercept and adds an inequality condition test. Algorithm 3 is modified accordingly to receive the regular observation and intercept point of each tree, and to calculate the path depth if the condition test is valid [40].

Algorithm 3: *PathLength* $(\vec{x}, T, e)$

Input: $\vec{x}$—an instance, T—an iTree, e—current path; to be initialized to zero when first called
Output: path length of $\vec{x}$
if T is an external node **then**
  return $e + c(T.size)\{c(.)$ is defined in Equation (8)$\}$
$\vec{m} \leftarrow T.Normal$
$\vec{p} \leftarrow T.Intercept$
if $(\vec{x} - \vec{p}).\vec{m} \leq 0$ **then**
  return $PathLength(\vec{x}, T.left, e + 1)$
else
  return $PathLength(\vec{x}, T.right, e + 1)$
end

Contamination is the parameters that estimate the number of outliers in a given set. The value is set near the confidence interval of 0.95, adapted for the Hotelling's distance-based models. Proposed values for the number of trees nt and the size of the ψ sub-sample are 100 and 256, respectively [34,35]. However, these parameters may vary according to the size and complexity of the dataset.

We carried out 50 simulations varying one parameter at a time while keeping the other fixed at its standard value. Figure 2 summarizes the results of these simulations, in which the points represent the average values calculated by the metric, while the bars represent the confidence interval of 0.95. The nt search parameter space is defined as 1, 5, 10, 50, 100, 500, and 1000, while ψ is the sample space following the power of 2, from 2^7 to 2^{13}. By varying the nt, we find that, with the increase in the number of trees, the variance and average of the TTC and CTT decrease, with the model showing excellent stability with 500 trees. Performing the same analysis for ψ, we found that, for values above 2048, CTT starts to increase gradually. Table 1 presents the parameters adopted for EIF based on these observations.



(**a**) (**b**)

Figure 2. Graphical representation for 25 simulations of the EIF model varying the control parameters of the algorithm. The points represent the average values and the bars the confidence interval of 0.95 calculated for the CTT and TTC metrics. In (**a**), the value of nt is changed and ψ is fixed at 256, while, in (**b**), the value of ψ is changed and nt is fixed at 100.

Table 1. Parameters adopted for iForest and EIF models.

Parameter	Value
Contamination	0.06
Number of trees—nt	500
Sub-sampling size—ψ	2048

2.5. Temporal Distance Metric

Whereas traditional anomaly detection adopts classification evaluation techniques, such as confusion matrix or one of its derived metrics [49], these metrics may be deficient in a time series context. Specific metrics, first developed for evaluating time series segmentation, are adopted in time series evaluation. In the present work, we adopt the average detection count and the absolute detection distance in order to evaluate the different methods [41].

Let $\mathbf{Y}$ be a time series of ordered set of real values indexed by natural numbers $\mathbf{Y} = \{y_0, y_1, y_2, ..., y_n\}, y_t \in \mathbb{R}$ and $\mathbf{C}$ a set of class labels or classification $\mathbf{C} = \{c_0, c_1, c_2, ..., c_n\}, c_t \in \{0, 1\}$, similar to $\mathbf{Y}$ but consisted of binary values $\{0, 1\}$. Values labeled as 0 represent normal values, and values labeled as 1 stand for anomalous values.

The average detection count l is simply given by the difference between the number of anomalies in the target classification and the candidate classification [41]:

$$l_{\downarrow}(\mathbf{C}_i) = |\text{count}(\mathbf{C}_0) - \text{count}(\mathbf{C}_i)| \tag{9}$$

where $\text{count}(\mathbf{C}_i) = \sum_{j=1}^{n} c_j, c_j \in C_i$.

The absolute temporal distance (TD) method consists of calculating the sum of all distances between anomalies from two classifications by [41]:

$$\text{TD}_{\downarrow}(C_i) = \text{TTC} + \text{CTT} \tag{10}$$

where TTC appears for Target To Candidate and is calculated by $\text{f}_{\text{closest}}(\mathbf{C}_0, \mathbf{C}_i)$, and CTT means Candidate To Target given by $\text{f}_{\text{closest}}(\mathbf{C}_i, \mathbf{C}_0)$. $\mathbf{C}_0$ and $\mathbf{C}_i$ denote target classification and candidate classification, respectively.

Note that lower values scored in each of the individual metrics are better than higher ones. Figure 3 graphically represents the calculation of the metric. In the given example, $\text{TTC} = \Delta t_1 + \Delta t_2$ and $\text{CTT} = \Delta t_1$, thus the absolute temporal distance $\text{TD} = 2\Delta t_1 + \Delta t_2$. The best possible value for the metric is zero, comprising a perfect detection system.

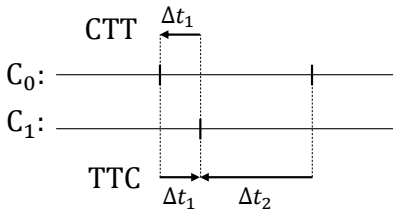


Figure 3. Temporal distance metric calculation. Adapted from [41].

2.6. Software and Hardware

The simulation was developed using the Python language, version 3.7.6, adopting common scientific libraries SciPy 1.4.1, Pandas 1.0.1 and NumPy 1.18.1. Scikit-learn 0.22.1 [50] presents efficient and reliable implementations of machine learning algorithms, such as PCA and Isolation forest. KICA-PCA was implemented by sequentially declaring the kernel approximation, ICA and PCA into a

pipeline of transformers. We applied the authors' original EIF algorithm implementation, available at the isotree 0.1.16 package [40].

Hardware specifications adopted to perform the simulation are: CPU Intel Core i7-8550U, 1.80 GHz, 16 GB RAM installed, and Windows 10 v.1909 operating system. The amount of time necessary to perform all 150 simulations is around 2 h. All data and scripts are available in the researcher's public repository (Github repository: https://github.com/rodrigosantis1/shp_anomaly).

3. Results and Discussion

The temporal distance and anomaly detection count are measured in test sets for each method, using Equations (9) and (10) shown in Section 2.5. The methods were trained 150 times using the training set, with unique random seeds. Table 2 exhibits the average and standard deviation of the calculated metrics. KICA-PCA obtained the smallest difference between real fault detection, in general. It was followed by PCA, EIF, and iForest. The methods with the lowest scores are shown in bold. As a linear model, PCA converges equally to the same solution when applied to a single training set. Thus, the standard deviation, unaffected by randomness, is not calculated for this particular method.

Table 2. Results of temporal distance and detection count obtained in simulations of anomaly detection models—standard deviation in parentheses and b.

	Temporal Distance (hours)			
Model	**TTC**	**CTT**	**TD**	**Detection Count—*l***
PCA	3270	738	4008	118
KICA-PCA	1633 (363)	933 (94)	2567 (375)	111.7 (15.7)
iForest	1541 (182)	934 (33)	2476 (185)	156.5 (6.1)
EIF	1474 (208)	906 (32)	2380 (210)	150.2 (5.0)

PCA is the method with the lowest CTT distance. This means that, since the distance between the detections and anomalies is the lowest, it is less likely to raise false alarms than the others. On the other hand, the TTC distance is higher with PCA than with all other methods. This means that the linear approach is less effective in detecting all anomalies. The total TD, obtained from the sum of the two individual components, is higher due to the TTC score.

Combining the PCA with the kernel trick and the ICA increased the accuracy, compared to the PCA alone. KICA-PCA presented intermediate distances when compared to all other methods. When compared to PCA, the nonlinear method improved the TTC, TD, and l. However, the main drawback is that its variance is higher than the others, meaning the method is more susceptible to randomness.

iForest presented the second-lowest TTC distance, which is the distance between the anomalies and the closest detection. Thus, this model was able to detect anomalies closest to their occurrence and, having the lowest TD and standard deviation, to present a suitable method for adoption in an online detection system. The main drawback with this method is the detection count: the number of detections is higher than the other methods. The EIF algorithm boosted the results obtained by iForest, with a TD reduction of 3.88% and an l reduction of 4.02% compared to its original implementation.

Figure 4 shows the anomaly score calculated using the EIF model for the entire test period. The moving average of the anomaly score illustrates that the observations are time-related, noting that the level of health tends to fluctuate over time. Factors like wear level, time of operation, maintenance, and shutdown directly impact the health index (HI).

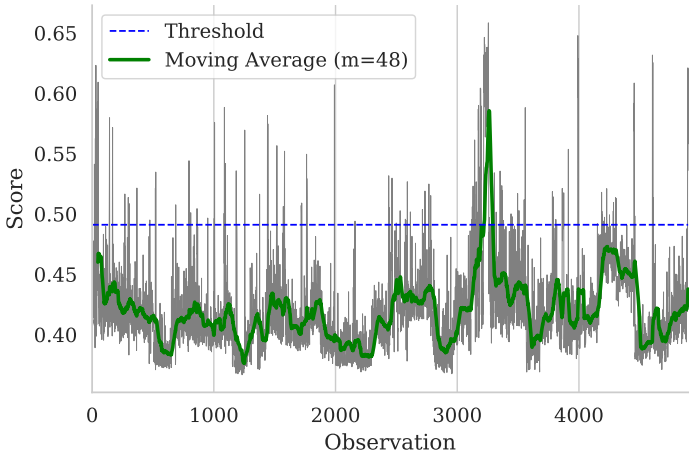


Figure 4. iForest anomaly score for the entire test period. The upper limit (threshold) represents the 0.95th quartile of the training data anomaly score and is shown in blue, while the moving average of 48 periods shows the trend of the series.

Analyzing some sample cases give us a better visualization of the model outcomes in practice. Figure 5 displays three selected time series excerpts, which comprise five faults out of the 59 registered in the dataset. iForest learned from the training set, and we analyzed the behavior of the model in a prominent failure situation. The *s* score represents the average count length of all iTrees. The threshold is set based on a confidence interval of 0.95. Every moment that the score exceeds the threshold is considered an anomaly. The actual anomalies are represented by red crosses, while the model detections are represented by blue dots. The observed excerpt periods are 17 December 2018, 18 January 2019, and 17 October 2018, respectively. Unconnected line periods represent missing data, which may occur when communication with the supervisory system is interrupted by network connection problems.

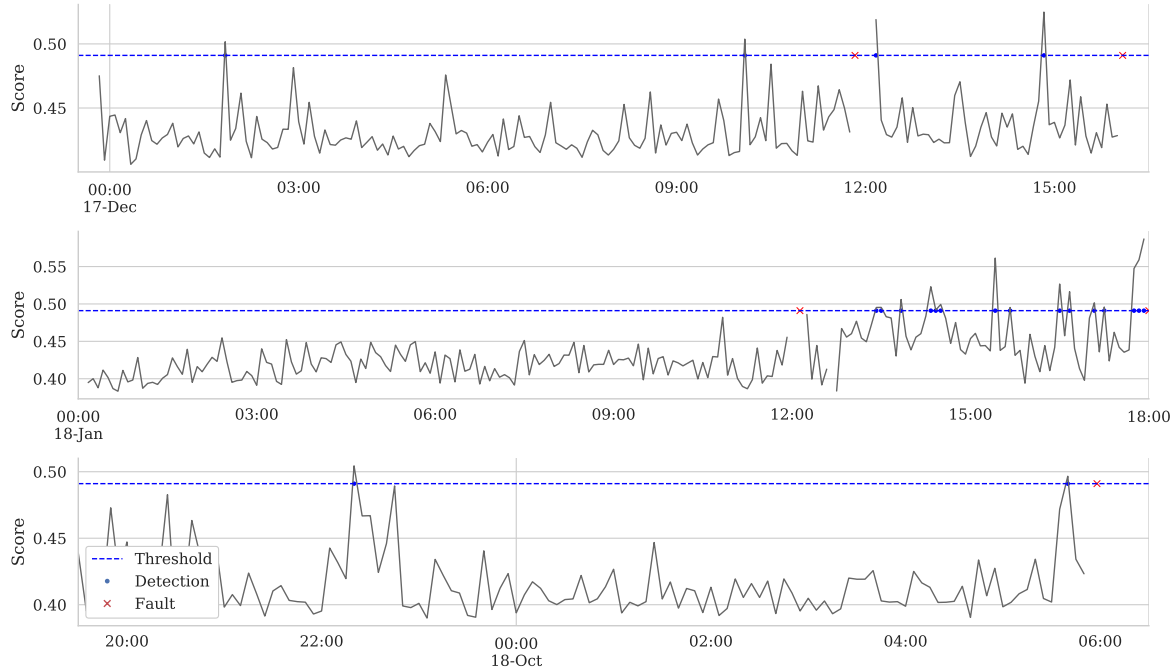


Figure 5. Three series excerpts, illustrating process monitoring time using EIF. Whenever the value calculated by the model exceeds the established limit, the measurement is considered an anomaly.

In the first period, the iForest model detected two anomalies before the first registered system failure. The first detection occurred 10 h before the failure, and the second detection occurred 2 h before

the failure. In this case, an intervention in the HGU system could have avoided the failure. A late detection was raised just after the first failure, which is considered a near detection. One hour before the second failure, the EIF model detected an anomaly. In total, the model detected four anomalies while two faults registered. The anomaly score profile calculated in this first excerpt is unsteady, with sudden fluctuations in the HI throughout the plotted period, making the detection task even more difficult.

The anomaly score increased over time in the second observed period, resulting in the first fault around noon. The HGU was rebooted and continued to operate, while iForest detected several anomalies. Six hours later, another fault occurred. In this example, the system could not detect any anomaly before the first failure, although it identified an abnormal state of operation between failures. It is possible to visually identify a positive growth trend in the anomaly score. The number of detections, in this example, associated with iForest results in Table 2, can explain the higher value of l and TTC.

In the third example, the anomaly score calculated between 10:00 p.m. and 11:00 p.m. detected an unusual operation, identifying two anomalies and presenting a situation in which the system nearly failed. However, the online monitoring system registered no fault until the next day. The model detects an anomaly minutes before the real fault. The damage is severe, and the system is shutdown. Throughout the rest of the observed period, the anomaly score was relatively low, close to 0.40.

Prognosis usually adopts the anomaly score to forecast the HGU behavior. This approach is commonly used to predict vibration trends and, since the score is a nonlinear combination of the monitored variables, it is likely that it also reflects the frequencies and amplitudes of the original signals. Some examples of vibration prediction models for prognosis can be found in [27,51].

The model has the limitation of not allowing the analysis of the importance of attributes, as do other models based on decision trees. The choice of the separation attributes in each node is random, and not generated from an explicit rule. However, machine understanding models, such as permutation-based and depth-based isolation forest, feature importance that can be used to circumvent this model limitation [52].

From our simulations, we found that EIF obtained an average TD reduction of 1628 (40.62%) compared to PCA, and of 187 (7.28%) compared to KICA-PCA. These results indicate that the anomaly detection algorithms are efficient and suitable for dealing with the problem of intelligent fault detection in hydroelectric plants, as indicated in the qualitative analysis of imminent failure. In some cases, the anomaly score depicts the trend in the risk of failure. In other cases, the anomaly score identifies regions of at-risk operation, even though no fault is registered.

Continuous improvement of the model is found in associating the detected fault patterns with known failure modes, using fault analysis techniques such as fault trees [53–55]. The anomaly score that is calculated can be used in future work to develop forecasting systems. The adoption of a single dimension HI simplifies the process control and the design of the predictive system. Instead of predicting each variable in isolation, one can focus on analyzing a single time series, which carries the individual characteristics of each of the individual measurement variables.

4. Conclusions

In the present paper, we propose the application of iForest for fault diagnosis in a small hydro-electric plant in CBM. The observed period is approximately one year, and the main input variables are vibration, oil inflow and apparent power. The model benchmarks, in the recently reported hydro-power fault diagnosis literature, are PCA and KICA-PCA, using the specific metrics of time series anomaly detection, temporal distance, and average detection count. The tree ensembles presented promising results, with lower error levels and variance than KICA-PCA. Another significant advantage of adopting iForest and EIF is their capability for parallel computing, which speeds up model training while keeping memory usage low, and fixed to a known limit.

Identifying failures before they occur is vital to allowing better management of asset maintenance, reducing operating costs and, in the case of SHPs, enabling the expansion of renewable energy sources in the energy matrix [56]. With the application of machine learning models such as iForest and EIF, the aim is to improve the health of the equipment and reduce power generation downtime.

Future studies should include investigating feature and model selection through exhaustive searching, Bayesian or evolutionary optimization, as parameters manually adjusted. Fine-tuning the models can contribute even more to increasing model accuracy. A step towards the prognostic model can be taken from the prediction of the anomaly score by decomposing the signal into components in the time and frequency spectrum, and combining methods of extracting attributes with uni- or multi-variate forecasting [28,51].

Another essential beneficial area of the present study is identifying feature importance in a SHP diagnosis system. This knowledge can guide the development of CBM systems by prioritizing the installation of critical sensors in SHP automation projects. EIF, since it is a generalization of iForest, can be combined with forward selection component analysis [57] for automatic variable selection.

Finally, the present study contributes to the improvement of SHP maintenance, a vital renewable power resource with huge potential for energy supply worldwide. By identifying faults before failure, management can take actions to avoid further damage caused to joint systems and further aggravation of the components, lowering the operating costs of power plants.

Author Contributions: Conceptualization, R.B.d.S. and M.A.C.; methodology, R.B.d.S. and M.A.C.; software, R.B.d.S.; validation, R.B.d.S., M.A.C.; formal analysis, M.A.C.; data curation, R.B.d.S.; writing—original draft preparation, R.B.d.S.; writing—review and editing, M.A.C.; visualization, R.B.d.S.; supervision, M.A.C.; project administration, M.A.C.; funding acquisition, M.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Brasil Energia Inteligente (BEI), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Grant No. 141777/2019-2, and Pro-Reitoria de Pesquisa (PRPq) da Universidade Federal de Minas Gerais.

Acknowledgments: The authors would like to thank Ado Popinhak for making available the failure monitoring database used in this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

CBM	Condition Based Maintenance
CTT	Candidate to Target
EIF	Extended Isolation Forest
HGU	Hydro-Generator Unit
HI	Health Index
HLU	Hydraulic Lubrication Unit
ICA	Independent Component Analysis
iForest	Isolation Forest
iTree	Isolation Tree
KICA-PCA	Kernel Independent Component and Principal Component Analysis
LS-SVM	Least Square and Support Vector Machine
PCA	Principal Component Analysis
SHP	Small Hydroelectric Plant
TD	Temporal Distance
TTC	Target to Candidate

Nomenclature

The following nomenclature is adopted in this manuscript:

$\mathbf{A}$	mixing component matrix
$\mathbf{C}$	classification vector
CTT	candidate to target distance
$\mathbf{D}$	diagonal matrix of eigenvalues
$\mathbf{E}$	residual matrix
$f_{closest}$	closest distance
F_{α}	F-distribution
h	average of path length
k	radial basis function
l	detection count
m	number of variables
$\vec{m}$	random number with m dimension
n	number of observations
nt	number of trees
$\vec{p}$	intercept
$\mathbf{P}$	loading matrix
$\mathbf{R}$	co-variance matrix
s	anomaly score
$\mathbf{S}$	independent component matrix
t	score
$\mathbf{T}$	score matrix
T^2	Hotelling's score vector
T^2_{α}	threshold
TD	temporal distance
TTC	target to candidate distance
$\vec{x}$	vector of observations
$\mathbf{X}$	matrix of observations
$\mathbf{X}'$	matrix of observations transformed
$\tilde{\mathbf{X}}'$	matrix of observations transformed and whitened
$\mathbf{Y}$	observations time series
z	low dimensional
α	degree of freedom
Δt	time difference
ψ	sub-sampling size
Φ	kernel's lifting

References

1. WEC. *World Energy Insights Brief*; Technical Report; World Energy Council: London, UK, 2019.
2. UNIDO. *World Small Hydropower Development Report 2016*; Technical Report; United Nations Industrial Development Organization: Vienna, Austria, 2016.
3. Ferreira, J.H.I.; Camacho, J.R.; Malagoli, J.A.; Júnior, S.C.G. Assessment of the potential of small hydropower development in Brazil. *Renew. Sustain. Energy Rev.* **2016**, *56*, 380–387. [CrossRef]
4. Dursun, B.; Gokcol, C. The role of hydroelectric power and contribution of small hydropower plants for sustainable development in Turkey. *Renew. Energy* **2011**, *36*, 1227–1235. [CrossRef]
5. Ohunakin, O.S.; Ojolo, S.J.; Ajayi, O.O. Small hydropower (SHP) development in Nigeria: An assessment. *Renew. Sustain. Energy Rev.* **2011**, *15*, 2006–2013. [CrossRef]
6. Kaunda, C.S.; Kimambo, C.Z.; Nielsen, T.K. Potential of Small-Scale Hydropower for Electricity Generation in Sub-Saharan Africa. *ISRN Renew. Energy* **2012**, *2012*, 1–15. [CrossRef]
7. Bhat, V.I.K.; Prakash, R. Life Cycle Analysis of Run-of River Small Hydro Power Plants in India. *Open Renew. Energy J.* **2014**, *1*, 11–16. [CrossRef]

8. Suwanit, W.; Gheewala, S.H. Life cycle assessment of mini-hydropower plants in Thailand. *Int. J. Life Cycle Assess.* **2011**, *16*, 849–858. [CrossRef]
9. Alonso-Tristán, C.; González-Peña, D.; Díez-Mediavilla, M.; Rodríguez-Amigo, M.; García-Calderón, T. Small hydropower plants in Spain: A case study. *Renew. Sustain. Energy Rev.* **2011**, *15*, 2729–2735. [CrossRef]
10. Kaldellis, J.K.; Vlachou, D.S.; Korbakis, G. Techno-economic evaluation of small hydro power plants in Greece: A complete sensitivity analysis. *Energy Policy* **2005**, *33*, 1969–1985. [CrossRef]
11. Bousdekis, A.; Magoutas, B.; Apostolou, D.; Mentzas, G. Review, analysis and synthesis of prognostic-based decision support methods for condition based maintenance. *J. Intell. Manuf.* **2018**, *29*, 1303–1316. [CrossRef]
12. Peng, Y.; Dong, M.; Zuo, M.J. Current status of machine prognostics in condition-based maintenance: A review. *Int. J. Adv. Manuf. Technol.* **2010**, *50*, 297–313. [CrossRef]
13. Sikorska, J.Z.; Hodkiewicz, M.; Ma, L. Prognostic modelling options for remaining useful life estimation by industry. *Mech. Syst. Signal Process.* **2011**, *25*, 1803–1836. [CrossRef]
14. Si, X.S.; Wang, W.; Hu, C.H.; Zhou, D.H. Remaining useful life estimation—A review on the statistical data driven approaches. *Eur. J. Oper. Res.* **2011**, *213*, 1–14. [CrossRef]
15. Salomon, C.P.; Ferreira, C.; Sant'Ana, W.C.; Lambert-Torres, G.; da Silva, L.E.B.; Bonaldi, E.L.; de Lacerda de Oliveira, L.E.; Torres, B.S. A study of fault diagnosis based on electrical signature analysis for synchronous generators predictive maintenance in bulk electric systems. *Energies* **2019**, *12*, 1506. [CrossRef]
16. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [CrossRef]
17. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
18. Ayo-Imoru, R.M.; Cilliers, A.C. Continuous machine learning for abnormality identification to aid condition-based maintenance in nuclear power plant. *Ann. Nucl. Energy* **2018**, *118*, 61–70. [CrossRef]
19. Li, H.; Chen, D.; Tolo, S.; Xu, B.; Patelli, E. Hamiltonian Formulation and Analysis for Transient Dynamics of Multi-Unit Hydropower System. *J. Comput. Nonlinear Dyn.* **2018**, *13*, 1–10. [CrossRef]
20. Wu, G.; Tong, J.; Zhang, L.; Zhao, Y.; Duan, Z. Framework for fault diagnosis with multi-source sensor nodes in nuclear power plants based on a Bayesian network. *Ann. Nucl. Energy* **2018**, *122*, 297–308. [CrossRef]
21. García Márquez, F.P.; Tobias, A.M.; Pinar Pérez, J.M.; Papaelias, M. Condition monitoring of wind turbines: Techniques and methods. *Renew. Energy* **2012**, *46*, 169–178. [CrossRef]
22. Tian, Z.; Jin, T.; Wu, B.; Ding, F. Condition based maintenance optimization for wind power generation systems under continuous monitoring. *Renew. Energy* **2011**, *36*, 1502–1509. [CrossRef]
23. Ding, H.; Ding, K.; Zhang, J.; Wang, Y.; Gao, L.; Li, Y.; Chen, F.; Shao, Z.; Lai, W. Local outlier factor-based fault detection and evaluation of photovoltaic system. *Sol. Energy* **2018**, *164*, 139–148. [CrossRef]
24. Kaid, I.; Hafaifa, A.; Guemana, M.; Hadroug, N.; Kouzou, A.; Mazouz, L. Photovoltaic system failure diagnosis based on adaptive neuro fuzzy inference approach: South Algeria solar power plant. *J. Clean. Prod.* **2018**, *204*, 169–182. [CrossRef]
25. Liu, X.; Kruger, U.; Littler, T.; Xie, L.; Wang, S. Moving window kernel PCA for adaptive monitoring of nonlinear processes. *Chemom. Intell. Lab. Syst.* **2009**, *96*, 132–143. [CrossRef]
26. Žvokelj, M.; Zupan, S.; Prebil, I. EEMD-based multiscale ICA method for slewing bearing fault detection and diagnosis. *J. Sound Vib.* **2016**, *370*, 394–423. [CrossRef]
27. Fu, W.; Wang, K.; Zhang, C.; Tan, J. A hybrid approach for measuring the vibrational trend of hydroelectric unit with enhanced multi-scale chaotic series analysis and optimized least squares support vector machine. *Trans. Inst. Meas. Control* **2019**, *41*, 4436–4449. [CrossRef]
28. Qiao, L.; Chen, Q. Forecasting Models for Hydropower Unit Stability Using LS-SVM. *Math. Probl. Eng.* **2015**, *2015*. [CrossRef]
29. Vu, V.H.; Thomas, M.; Lafleur, F.; Marcouiller, L. Towards an automatic spectral and modal identification from operational modal analysis. *J. Sound Vib.* **2013**, *332*, 213–227. [CrossRef]
30. Peng, W.J.; Luo, X.Q.; Guo, P.C.; Lu, P. Vibration fault diagnosis of hydroelectric unit based on LS-SVM and information fusion technology. *Zhongguo Dianji Gongcheng Xuebao/Proc. Chin. Soc. Electr. Eng.* **2007**, *27*, 86–92. [CrossRef]
31. Gregg, S.W.; Steele, J.P.; Van Bossuyt, D.L. Feature selection for monitoring erosive cavitation on a hydroturbine. *Int. J. Progn. Health Manag.* **2017**, *8*, 1–22.

32. Ge, Z.; Song, Z. Process monitoring based on independent Component Analysis-Principal Component Analysis (ICA-PCA) and similarity factors. *Ind. Eng. Chem. Res.* **2007**, *46*, 2054–2063. [CrossRef]
33. Zhu, W.; Zhou, J.; Xia, X.; Li, C.; Xiao, J.; Xiao, H.; Zhang, X. A novel KICA-PCA fault detection model for condition process of hydroelectric generating unit. *Meas. J. Int. Meas. Confed.* **2014**, *58*, 197–206. [CrossRef]
34. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data* **2012**, *6*. [CrossRef]
35. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
36. Sun, L.; Versteeg, S.; Boztas, S.; Rao, A. Detecting Anomalous User Behavior Using an Extended Isolation Forest Algorithm: An Enterprise Case Study. *arXiv* **2016**, arXiv:1609.06676.
37. Susto, G.A.; Beghi, A.; McLoone, S. Anomaly Detection through online Isolation Forest: An application to plasma etching. In Proceedings of the 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), Saratoga Springs, NY, USA, 15–18 May 2017; pp. 89–94. [CrossRef]
38. Sureda Riera, T.; Bermejo Higuera, J.R.; Bermejo Higuera, J.; Martínez Herraiz, J.J.; Sicilia Montalvo, J.A. Prevention and Fighting against Web Attacks through Anomaly Detection Technology. A Systematic Review. *Sustainability* **2020**, *12*, 4945. [CrossRef]
39. Vartouni, A.M.; Kashi, S.S.; Teshnehlab, M. An anomaly detection method to detect web attacks using Stacked Auto-Encoder. In Proceedings of the 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems, CFIS 2018, Kerman, Iran, 28 February–2 March 2018; pp. 131–134. [CrossRef]
40. Hariri, S.; Kind, M.C.; Brunner, R.J. Extended Isolation Forest. *IEEE Trans. Knowl. Data Eng.* **2019**. [CrossRef].
41. Kovács, G.; Sebestyen, G.; Hangan, A. Evaluation metrics for anomaly detection algorithms in time-series. *Acta Univ. Sapientiae Inform.* **2020**, *11*, 113–130. [CrossRef]
42. Hempstalk, K.; Frank, E.; Witten, I.H. One-class classification by combining density and class probability estimation. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Antwerp, Belgium, 15–19 September 2008; pp. 505–519.
43. Khan, S.S.; Madden, M.G. A survey of recent trends in one class classification. In Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, 19–21 August 2009; pp. 188–197.
44. Liao, G.P.; Gao, W.; Yang, G.J.; Guo, M.F. Hydroelectric Generating Unit Fault Diagnosis Using 1D Convolutional Neural Network and Gated Recurrent Unit in Small Hydro. *IEEE Sens. J.* **2019**, *19*, 9352–9363. [CrossRef]
45. Navi, M.; Meskin, N.; Davoodi, M. Sensor fault detection and isolation of an industrial gas turbine using partial adaptive KPCA. *J. Process Control* **2018**, *64*, 37–48. [CrossRef]
46. Rahimi, A.; Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20—Proceedings of the 2007 Conference*; Curran Associates Inc.: Reed Hook, NY, USA, 2009; pp. 1–8.
47. Ring, M.; Eskofier, B.M. An approximation of the Gaussian RBF kernel for efficient classification with SVMs. *Pattern Recognit. Lett.* **2016**, *84*, 1339–1351. [CrossRef]
48. Senechal, T.; McDuff, D.; El Kaliouby, R. Facial Action Unit Detection Using Active Learning and an Efficient Nonlinear Kernel Approximation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 10–18. [CrossRef]
49. Zemouri, R.; Levesque, M.; Amyot, N.; Hudon, C.; Kokoko, O.; Tahan, S.A. Deep Convolutional Variational Autoencoder as a 2D-Visualization Tool for Partial Discharge Source Classification in Hydrogenerators. *IEEE Access* **2020**, *8*, 5438–5454. [CrossRef]
50. Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Pedregosa, F.; Mueller, A. Scikit-learn. *GetMobile Mob. Comput. Commun.* **2015**, *19*, 29–33. [CrossRef]
51. Zhou, K.B.; Zhang, J.Y.; Shan, Y.; Ge, M.F.; Ge, Z.Y.; Cao, G.N. A hybrid multi-objective optimization model for vibration tendency prediction of hydropower generators. *Sensors* **2019**, *19*, 2055, [CrossRef] [PubMed]
52. Carletti, M.; Masiero, C.; Beghi, A.; Susto, G.A. Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 21–26. [CrossRef]
53. Jong, C.G.; Leu, S.S. Bayesian-network-based hydro-power fault diagnosis system development by fault tree transformation. *J. Mar. Sci. Technol. (Taiwan)* **2013**, *21*, 367–379. [CrossRef]

54. Melani, A.H.; Silva, J.M.; de Souza, G.F.; Silva, J.R. Fault diagnosis based on Petri Nets: the case study of a hydropower plant. *IFAC-PapersOnLine* **2016**, *49*, 1–6. [CrossRef]
55. Cheng, J.; Zhu, C.; Fu, W.; Wang, C.; Sun, J. An Imitation medical diagnosis method of hydro-turbine generating unit based on Bayesian network. *Trans. Instit. Meas. Control* **2019**, *41*, 3406–3420. [CrossRef]
56. Zhang, Y.; Zhao, X.; Zuo, Y.; Ren, L.; Wang, L. The development of the renewable energy power industry under feed-in tariff and renewable portfolio standard: A case study of China's photovoltaic power industry. *Sustainability* **2017**, *9*, 532. [CrossRef]
57. Puggini, L.; McLoone, S. An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. *Eng. Appl. Artif. Intell.* **2018**, *67*, 126–135. [CrossRef]