

Article

Condition-Based Maintenance of Gensets in District Heating Using Unsupervised Normal Behavior Models Applied on SCADA Data

Valerio Francesco Barnabei ^{1,*} , Fabrizio Bonacina ¹ , Alessandro Corsini ¹ , Francesco Aldo Tucci ¹ and Roberto Santilli ²

¹ Department of Mechanical and Aerospace Engineering, University of Rome La Sapienza, Via Eudossiana 18, I00184 Rome, Italy

² ENGIE Servizi S.p.A, District Heating and Power, Viale Avignone 12, I00144 Rome, Italy

* Correspondence: valerio.barnabei@uniroma1.it

Abstract: Increasing interest in natural gas-fired gensets is motivated by District Heating (DH) network applications, especially in urban areas. Even if they represent customary solutions, when used in DH, duty regimes are driven by network thermal energy demands resulting in discontinuous operation, which affects their remaining useful life. As such, the attention on effective condition-based maintenance has gained momentum. In this paper, a novel unsupervised anomaly detection framework is proposed for gensets in DH networks based on Supervisory Control And Data Acquisition (SCADA) data. The framework relies on multivariate Machine-Learning (ML) regression models trained with a Leave-One-Out Cross-Validation method. Model residuals generated during the testing phase are then post-processed with a sliding threshold approach based on a rolling average. This methodology is tested against nine major failures that occurred on the gas genset installed in the Aosta DH plant in Italy. The results show that the proposed framework successfully detects anomalies and anticipates SCADA alarms related to unscheduled downtime.

Keywords: multivariate time series; early fault detection; condition based maintenance; multi-MW gensets SCADA data



Citation: Barnabei, V.F.; Bonacina, F.; Corsini, A.; Tucci, F.A.; Santilli, R. Condition-Based Maintenance of Gensets in District Heating Using Unsupervised Normal Behavior Models Applied on SCADA Data. *Energies* **2023**, *16*, 3719. <https://doi.org/10.3390/en16093719>

Academic Editors: Shunli Wang, Jiale Xie and Guang Wang

Received: 14 March 2023

Revised: 14 April 2023

Accepted: 19 April 2023

Published: 26 April 2023



Copyright: © 2023 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

District Heating, also known as heat networks or teleheating, provides a platform for heat supply based on the integration of low-carbon technologies, including renewable energy sources and thermal storage, to improve overall efficiency and minimize greenhouse gas emissions. In operation since the end of the XIX century, DH represents an efficient way to provide heat to a large number of users in densely populated urban areas [1–5]. According to IEA’s 2021 report [6], DH systems are important solutions to describe the heating sector in any NZE 2050 scenario [7].

DH systems are composed of thermal plants and a distribution network of insulated pipes that deliver heat to the end users. The thermal plant is based on technology to generate heat from fossil fuels or renewable energy sources or to valorize waste heat [8]. In 2020, nearly 90% of heat was produced from fossil fuels, and one of the most common technologies in DH thermal power plants involves the use of generator sets, also known as gensets, with internal combustion engines (ICEs) either in combined heat and power (CHP) configurations or directly coupled with heat pumps [9].

Wang et al. [10] reported that, in 2012, in China, more than 36% of the total building energy demand was consumed for residential heating purposes, and about 62.9% of district heat was produced by CHP systems. As another example, in Finland, DH accounts for about 50% of the total heating market, and the city of Helsinki has around 20% of their district heat produced by genset with the use of wastewater as a low-grade heat source [11].

Gensets can suffer from intermittent operation caused by the variability and seasonality of the network heat demand, especially when directly coupled with heat pumps. These operation modes often lead the engine off-design and can be interpreted as the root cause of genset anomalies and failures. Therefore, the research on automatic Fault Detection (FD) of gensets based on proper Condition-Based Maintenance (CBM) strategies is of paramount importance to monitor the operation, reduce downtime and ensure the reliability and productivity of the overall heat supply process [12–14].

Rooted in condition-monitoring systems, CBM aims to establish frameworks for the diagnosis of equipment under supervision indicating incipient failures using sensor networks. CBM defines and monitors health indicators capable of signaling an anomaly in the case of deviation from reference values. Based on the evaluation of the current state of the equipment, it is possible to identify faults and malfunctions at an early stage, thus, allowing the timely planning of maintenance interventions. Despite the fact that scheduled maintenance and CBM are complementary, CBM is, by far, the most cost-effective approach and the one that enhances the life expectancy of the equipment [15,16].

A recent review on ICE diagnostics [17] suggested that a limited number of papers dealt with analytical models specifically designed for the CBM of gensets operating in DH networks. Most of the literature is dedicated to load prediction and the analysis of optimal network design with few contributions focusing on the operation and maintenance of networks and distribution pipelines [18].

As reported in [19], Machine-Learning (ML) algorithms have also been established as a viable solution in the DH scenario because they are easily adaptable to changing conditions, capable of modeling non-linear phenomena and can benefit from the historical data readily available in modern control systems (e.g., SCADA data). While ML approaches based on classification algorithms, such as Bayesian Classifiers (BCs) or Support Vector Machines (SVMs), have been widely used for FD of ICEs [20–25], regression algorithms seem to represent the most suitable option to perform an effective CBM.

In fact, on the one hand, BCs and SVMs are supervised ML tools that enable effective FD, but they rely on events that already occurred in the past to label the training dataset. On the other hand, unsupervised models based on regression approaches, classified in [26] as Normal Behavior Models (NBM), are able to detect anomalies in real-time conditions, as they can signal upcoming fault events in advance.

As a general outline, NBM approaches for CBM consist of training a reference model that represents the normal operation of the system and evaluating the deviation, or residual, between the predicted and actually measured values in real-time conditions to detect anomaly occurrence. Note that training a regression model to create an NBM may appear to be a supervised approach because it is trained on examples in which the expected values of the target variable are also provided; however, due to the absence of labels classifying the operational state in the training phase, NBM models fall into the category of unsupervised fault-detection methods [27].

The scope of this work is to propose an unsupervised NBM model designed for gensets operating in DH networks that introduces a series of advantages with respect to the state of art as detailed in the following section.

2. Unsupervised CBM of ICEs: State of the Art

To date, most applications of data-driven unsupervised fault detection in ICE fall in the automotive, aviation and marine sectors. To name a few, Liu et al. [28] used a linear regression based on thermal and electrical parameters for detecting the valve clearance of diesel engines. Bryght et al. [29] predicted failure in aircraft engines by combining lead function and logistic regression applied to aircraft engine takeoff data. Singh et al. [30] tested the performance of several Machine-Learning algorithms for predicting the health of an aircraft engine on historical data retrieved from the NASA data repository. Maraini et al. [31] developed a data-driven framework based on a Multi-Layer Perceptron (MLP) for marine

gas turbine engine health monitoring. Chen et al. [32] proposed a deep autoencoder with a Dimension Fusion Function method (DFF-DAE) to detect aero-engine faults.

Focusing on the specific applications of ICEs in power plants, Mendonça et al. [12] proposed a methodology for the detection of incipient failures in the components of internal combustion engine-driven generators based on Electrical Signature Analysis (ESA), while Deon et al. [33] introduced a predictive maintenance module within a digital twin based on the definition of independent subsystems, each one supported by an ad hoc trained model (Air Intake Subsystem, Exhaust Subsystem, Fuel Subsystem, Water Cooling Subsystem, Lubrication Subsystem and Mechanical Subsystem).

Based on the above, it can be concluded that a large part of the literature envisages the development of different Machine-Learning models applied to data sampled from sensor networks specifically designed for condition-monitoring systems (e.g., accelerometers and vibration sensors). On the other hand, as an interesting perspective, in recent years, there has been an increasing focus on NBM approaches for CBM based on SCADA data, especially in the context of wind turbines (see [26] for a comprehensive review).

However, NBM approaches can present a number of critical issues when applied to multivariate SCADA data. In this sense, a number of challenges were identified in [27]. As a first example, the high data dimensionality heavily affects the response times of NBM models, making them frequently unsuitable for near real-time applications typical of CBM. A second concern is represented by the challenge in isolating the size of the time window to train the reference model: the seasonal nature of the operating conditions, coupled with the possible presence of undesired anomalies in the dataset, makes it difficult to identify the standard dynamics of the system using, for example, standard approaches for clustering or outlier isolation.

Finally, a further issue is represented by the appropriate handling of residuals for alarm activation. Since residuals are evaluated as the difference between the value of a signal predicted by the regression model (trained under reference conditions) and the actual value of the same signal logged by the SCADA sensor, they can present a high level of noise and typical signal variability, which makes it very challenging to trigger alarms using standard control charts.

As an attempt to face the aforementioned issues and challenges, a general framework for SCADA-based CBM using a NBM approach is proposed, and the method is applied to the technology of natural gas (NG) gensets in DH networks. Specifically, the framework proposes a series of solutions to manage the entire data-mining process, starting from the reduction of dimensionality in the pre-processing phase with a feature-selection algorithm, passing through the training methods of the reference models with a Leave-One-Out Cross-Validation approach [34], up to the post-processing of residuals by means of the introduction of a two-stage sliding threshold metric to provide nowcasting of the alarms. For the ML module, two different regression algorithms, namely, XGB and MLP, are trained and compared.

The framework is tested on SCADA data sampled on a 7.5 MW NG genset installed in the District Heating plant of the city of Aosta, Italy. The considered dataset includes 45 parameters with 5 min sampling during 16 months of engine operation (from September 2019 to December 2020). The paper is organized as follows. Section 3 presents the discussion of the building blocks of the proposed ML framework for CBM. Then, Section 4 describes the case study and the obtained results. Finally, Section 5 summarizes the present work and presents our conclusions.

3. Anomaly Detection Framework, Overview

The first operation proposed in the framework is the pre-processing and cleaning of SCADA event logs and signals, filtering out minor events from the logs and removing constant signals (see Section 3.1).

Subsequently, we process all SCADA signals with a feature-selection method based on a variable importance approach to select the best predictors for the nowcasting of a specific

target variable (see Section 3.2). These preliminary operations optimize the performance of the ML models both in terms of accuracy and computational costs for CBM purposes.

In the next step, we apply two completely different models (namely, XGB and MLP) independently for the construction of the reference model, training both of them with a Leave-One-Out Cross-Validation approach (see Section 3.3). This avoids any risk of overfitting and guarantees greater robustness and flexibility of the results by simulating unsupervised real-time applications. We recommend having at least one year of data for the training phase, to guarantee the effective learning of the recurring relational dynamics between signals while still taking into account the seasonal operational variations typical of the analyzed users.

At the testing stage, we adopt a warning rule for anomaly detection based on a sliding threshold metric approach, applied to the Local Residual Indicators (LRIs) of each parameter. Specifically, we filter the noise of LRIs and subsequently define a control chart based on their intensity and time persistence to trigger alarms only related to significant anomalies and to reduce the occurrence of false positives (see Section 3.5).

Finally, we evaluate the anomaly detection results with respect to the ability to identify precursors from the SCADA event logs and early detect major faults. Concerning the SCADA event logs, after a preliminary filtering of minor events, the framework integrates the evaluation of the Mean Time Between Alarms (MTBA) indicator and the quantification of the total downtime in a prognostic perspective.

The entire framework is implemented using Python 3.9 Scikit-Learn open-source library [35]. A step-by-step framework description is given in the following Figure 1.

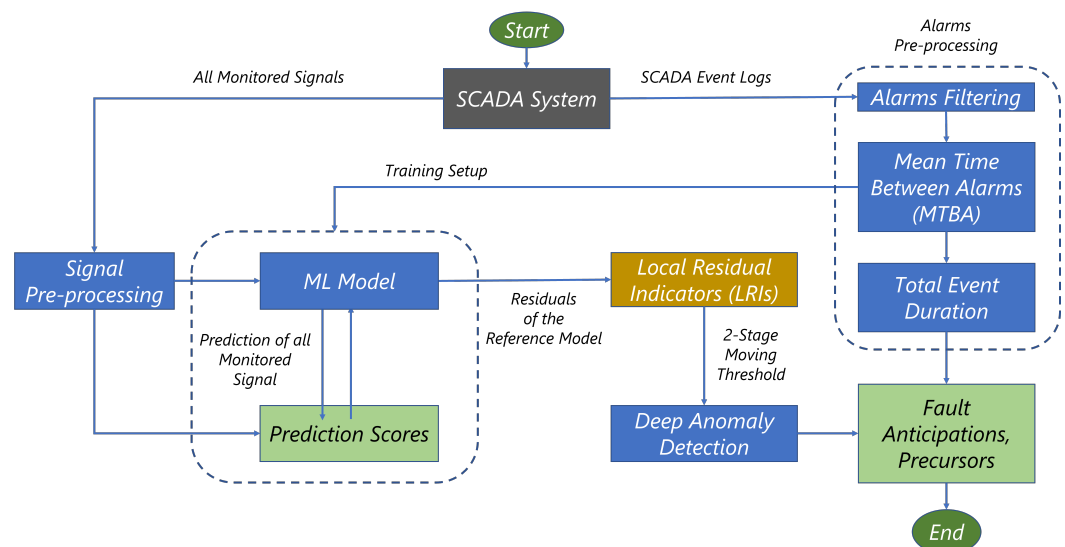


Figure 1. ML framework for CBM and schematics.

3.1. SCADA Event Log and Signal Pre-Processing

The pre-processing of the SCADA event logs filters all minor alarms unrelated to specific faults or anomalies, along with events recorded during the engine downtime. The remaining logs are then used to estimate operation metrics, such as the MTBA and the total duration of the outage events until correct operations are recorded. Those indicators represent key parameters for the training setup of the ML model (as explained in more detail in Section 3.4). Additionally, we evaluated the information content of each signal time series using the Shannon Entropy (H) metric [36], which allows for the interpretation of parameters with H close to zero as irrelevant or derived and to remove them from the training dataset, together with constant signals. Finally, a sigma rule was adopted to identify and remove extreme outliers related to measurement errors and to finally filter the signals with respect to the active power of the ICE.

3.2. Feature Selection

The framework adopts a feature-selection method based on variable importance through exploiting the Predictive Power Score (PPS) [37] algorithm. The output of the PPS analysis is an asymmetric, data-type independent index that identifies the relationships among the features in a dataset. Specifically, PPS quantifies how much a single input variable affects the prediction of the target variable. PPS assigns an index on each single input feature (x_i) at a time used to predict the target variable (y_i) via a Decision Tree algorithm. The index is expressed as:

$$PPS = 1 - \frac{MAE_{model}^{x_i, y_i}}{MAE_{naive}^{y_i}} \quad (1)$$

where $MAE_{model}^{x_i, y_i}$ is the Mean Absolute Error of the chosen regression model that predicts y_i from a candidate x_i , while $MAE_{naive}^{y_i}$ is obtained with a naive model that always predicts the median of y_i . The index ranges from 0 (no predictive power) to 1 (perfect predictive power). On this basis, as suggested by the authors of the algorithm [37], the minimum PPS acceptability limit is consistently set at 0.2. For each specific target variable (y_i), a vector of best predictors B_i is defined, selecting from the set of all possible input features (x_i), the ones with a PPS score above the set threshold. For example, as highlighted in Figure 2, for the specific target variable (y_i) the vector of best predictors B_i includes the subset of input features ranging from x_1 to x_8 .

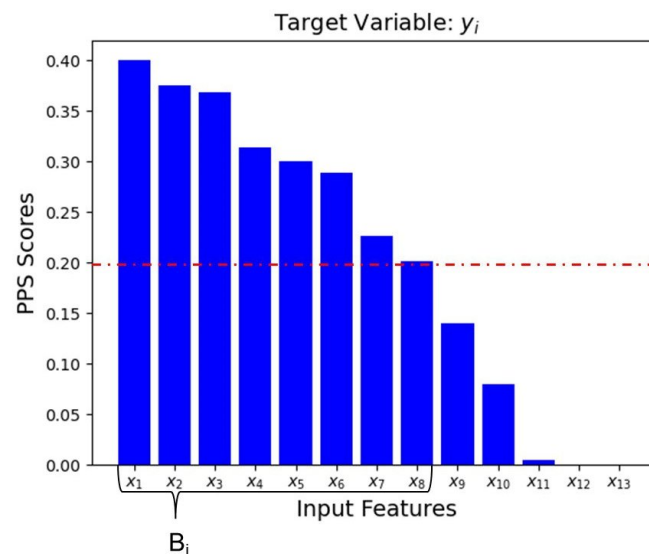


Figure 2. Example of criteria used to select best predictors based on the PPS score. Bars represent the score value, while the red dashed line represent the minimum acceptability limity for the score.

3.3. Machine-Learning Model

Two different regression algorithms, namely, XGB and MLP, are selected as candidates for the ML module. Both the regression algorithms saw an optimization process using a grid search approach [38] to select the best combination of hyper-parameters. In spite of the fact that both models identify within the training dataset one parameter at a time as the target variable (y_i) and exploit all the others to predict it, some core differences between the models still represent a challenge for comparability.

Notably, since XGB belongs to the category of ensemble algorithms and since its structure is composed of several decision trees, the results are independent from feature normalization [39]. In contrast, Artificial Neural Networks rely on statistical analysis and, thus, are strongly influenced by the distribution and quality of the data and are highly dependent on the order of magnitude of their input values. As a consequence, MLP may

neglect or overestimate the influence of certain features according to their values [40]. To avoid this, input signals are initially normalized for the MLP model using a Standard Scaler and then the predicted features are scaled back to their original size. This ensures the comparability of results between the two ML models in terms of prediction scores.

3.4. Training Setup

As previously stated, the training strategy relies on a Leave-One-Out Cross-Validation method [34] as a proposed solution to isolate reference operating conditions with standard unsupervised approaches in highly discontinuous duty periods combined with the strong seasonality of the signals. In the specific DH application presented in the paper, the genset workload presented strong discontinuities in the summer period as well as a higher environmental temperature operating condition, while having a more continuous workload in winter with lower external temperatures.

In detail, as shown in Figure 3, one month m is cyclically isolated as the testing dataset D_{test} , and a model is trained on the remaining months split between training D_{train} and validation D_{val} datasets. This approach is meant to avoid possible overfitting and presume that most of the operational data over a long period of time refers to normal engine operation. To further reduce the possible presence of failure precursors in the reference model, D_{train} does not include any downtime period, considering an additional safety time range equal to the value of the MTBA index obtained at the pre-processing stage.

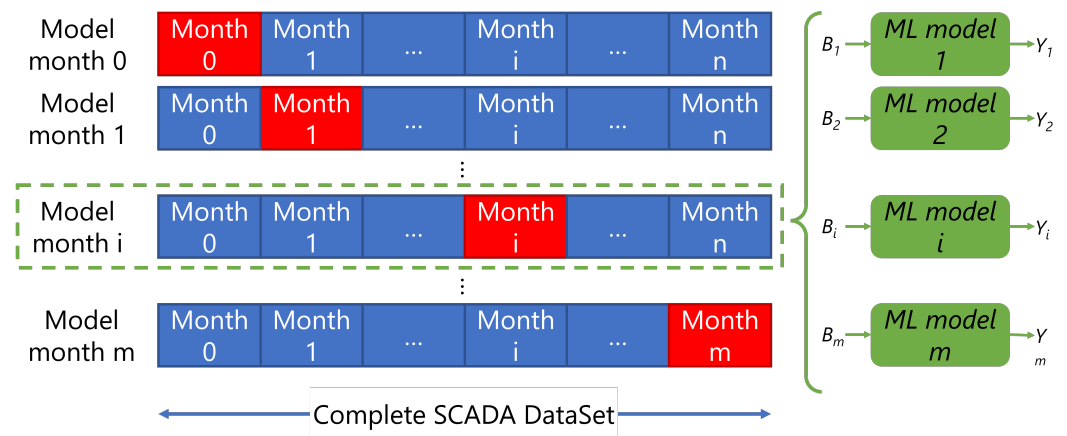


Figure 3. Representation of the Leave-One-Out Cross-Validation method as implemented in the present study.

As a result of this training process, a specific regression model (ML model $_i$) for each target variable (y_i) is obtained and defined as a function of best predictors B_i previously identified. The accuracy of the two models during the training phase on the reference period was evaluated with customary scores, i.e., the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MDAPE).

3.5. Residual Indicator Definition

The aim of the proposed CBM framework is the definition of anomaly detection rules to trigger early warnings of incipient failures. To this end, an *LRI* is defined for each monitored variable [41] as the absolute value of the difference between the actual values (f) and those predicted by the models trained on the reference period (f_p):

$$LRI = |f - f_p| \tag{2}$$

Additionally, the *LRI* is enhanced with a sliding threshold metric based on an average obtained with a rolling-window algorithm. This is done to trigger early warnings while limiting the occurrence of false alarms due to *LRI* spikes. In particular, as shown in Figure 4,

an alarm is triggered for a signal when the following condition is satisfied for P consecutive time steps:

$$LRI_i \geq 0.5 \cdot \frac{1}{W} \sum_{j=i-W}^i LRI_j \quad (3)$$

where LRI_i is the LRI of the signal at time i and W is the length of the sliding window. The value of W should be selected according to the periodicity of the observed phenomena and, in this specific case, corresponds to 24 h. Thus, as the averaged LRI experiences a deviation $\geq 50\%$ compared to the last 24 h that persists for at least P time steps, an alarm is triggered for the specific sensor of that LRI . We set the persistence threshold P to 6 h, which resulted in effectively removing residual noise.

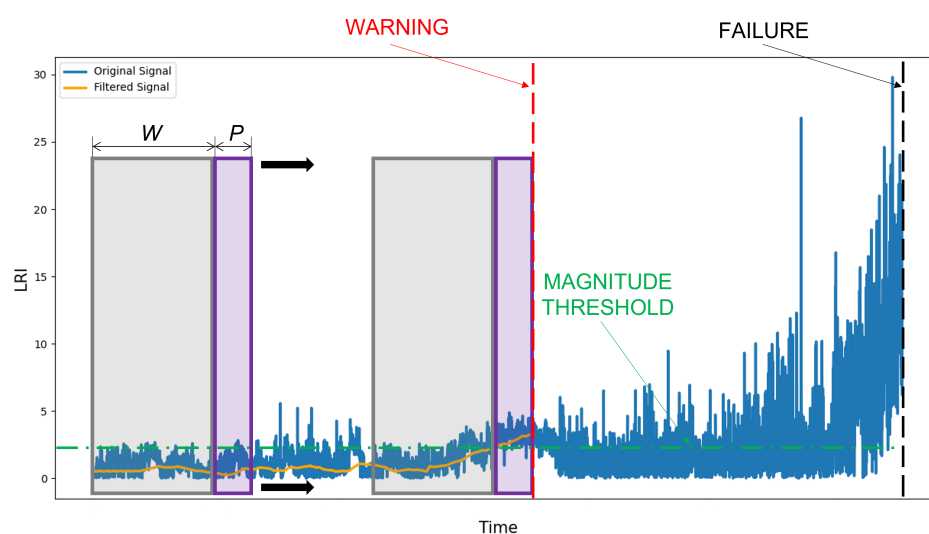


Figure 4. Example of the anomaly detection rule to trigger early alarms on specific sensors.

This approach proved to be particularly suitable for this type of dataset, in which a standard control chart with a fixed threshold for LRIs could be ineffective due to the extreme data variability in some periods and seasons. Moreover, it guarantees high robustness in handling the noise of the residuals of the models.

Based on such a warning rule, model performances were evaluated in terms of anomaly detection capability on each cross-validation datasets cyclically isolated. This assessment aims to quantify the ability of each warning to anticipate the major failure events included in the SCADA log.

4. Results

4.1. Dataset Description

Data were collected from a natural gas genset installed in the Aosta District Heating plant, which is equipped with a 16-cylinders turbocharged ICE. The engine has a nominal electric power output of 7.5 MWe, and it is directly coupled to a 17.5 MWt heat pump. ICE technical specifications are given in Table 1.

A SCADA system monitors different operating parameters collected by the main components of the genset together with environmental measurements. In detail, the initial dataset included 45 parameters sampled every 5 min from September 2019 to December 2020, for a total of 15 months. After the application of the signal preprocessing described in the Section 3.1, the feature number was reduced to 33 significant parameters as listed in Table 2.

Table 1. Technical specifications of the engine.

Quantity	Value	Unit
N. of cylinders	16	[-]
Engine speed	720	[r/min]
Electrical power output	7235	[kW]
Thermal power air cooler HT	1305	[kW]
Thermal power air cooler LT	490	[kW]
Thermal power lube oil cooler	730	[kW]
Thermal power jacket water cooler	925	[kW]
Exh. mass flow rate	39600	[kg/h]
Exh. gas temp.	355	[°C]

Table 2. List of SCADA signals.

Signal ID	Description
P01–P19, P23, P25–P26	Cylinder, exhaust and intake temperatures
P20–P22, P24	Cylinder and fuel subsystem pressures
P27–P31	Generator phase and bearing temperatures
P32	Active power
P33	Ambient temperature

In addition to the SCADA signals, the framework's anomaly detection capability was evaluated by looking at the alerts logged by the SCADA system from October to December 2020, a period when numerous major failures occurred. As described in Section 3.1, only major events were considered, including scheduled (i.e., normal stop) and unscheduled downtime (i.e., emergency stop or outages after engine deratings). Table 3 lists the filtered major SCADA events in the reference period.

Table 3. Event log for major events in the observation period. Event types are abbreviated as follows: D—Derating, NS—Normal Stop and UD—Unscheduled Downtime.

Event ID	SCADA Event Log	Event Type	Start	Duration (hh)
DS_05_10	Exh Temp Deviation Cylinder	D & UD	5 October 2020	11
S_07_10	Emergency Stop Activated	UD	7 October 2020	123
DS_15_10	Exh Temp Deviation Cylinder	D & UD	15 October 2020	11
S_20_10	Emergency Stop Activated	UD	20 October 2020	5
DS_26_10	Exh Temp Deviation Cylinder	D & UD	26 October 2020	11
S_13_11	Emergency Stop Activated	UD	13 November 2020	4
D_16_11	Charge Air Temp After Cooler High	D	16 November 2020	1
S_19_11	Emergency Stop Activated	UD	19 November 2020	48
S_13_12	Shutdown from Main Control	NS	13 December 2020	1
DS_16_12	Generator Stator Temp Windings	D & UD	16 December 2020	1
S_21_12	Emergency stop Activated	UD	21 December 2020	12

4.2. ML Settings and Prediction Errors

Both ML approaches experienced identical training, cross-validation and testing phases. At the training stage, the dataset was split into training and validation sets, respectively, named D_{train} and D_{val} , corresponding to 70% and 30% of the total set. Finally, the testing set D_{test} consisted of a single month cyclically isolated from the available data and included the time periods of failure occurrences.

The XGB model learning task was set to linear regression with hyperparameters optimization according to grid search algorithm, while the MLP setup included early stopping to avoid overfitting. Tables 4 and 5 lists the two subsets of hyperparameters.

Table 4. XGB regressor hyperparameters.

Hyperparameter	Value
Subsampling of columns	0.20
Learning rate	0.10
Max depth	50
Nr. of trees	150
Nr. of parallel trees	20
Alpha	0
Lambda	1

Table 5. MLP regressor hyperparameters.

Hyperparameter	Value
Nr. of Neurons	22
Nr. of hidden layer	1
Nr. of training epochs	150
Activation function	relu
Initial learning rate	1×10^{-5}
Optimizer	ADAM
Batch size	1/50th

The ML model predictions are evaluated in terms of the reconstruction errors of all SCADA signals (during the training phase the predicted values are compared with the actual ones). As can be seen from Table 6, XGB outperforms MLP in terms of customary scores.

Table 6. Reconstruction errors for the proposed ML models.

	XGB	MLP
MAE	0.04	0.11
MSE	0.10	0.14
RMSE	0.21	0.31
MDAPE	0.01	0.13

4.3. Anomaly Detection Results

For the evaluation of the anomaly detection capabilities, the results of the testing phase refers to the period of October to December 2020. Specifically, ML model results are discussed by plotting the LRIs against the relative warnings activated on the individual parameters after the application of the sliding threshold metrics (Equation (3)). Furthermore, as a reference to identify engine derating and shutdown, the results are presented in terms of the active power together with the details of the main alarms recorded by the SCADA system in the same time interval.

Figures 5–7 illustrate the results in October 2020. Figure 5 shows the active power, with the detail of SCADA event logs recorded in that period (event IDs refer to Table 3). Figures 6 and 7 show the LRI together with the warnings triggered by the framework (highlighted in dashed red lines).

By analyzing October 2020 SCADA logs, five significant events were isolated. Those events include three anomalies that resulted in a preliminary power output derating followed by engine shutdown, along with two emergency stops linked to unscheduled downtimes. Regarding the first event category, it is worth noting that all the shutdowns were anticipated by cylinder temperature anomalies and that the application of the proposed framework allows for the early detection of such precursors. In particular, for the events detected on 5 October 2020 (event ID: DS_05_10) and 15 October 2020 (event ID: DS_15_10), respectively, a significant deviation of the LRI associated with cylinder temperature parameter (P16) can be seen in Figures 6 and 7, resulting in early warnings with

respect to the actual SCADA log (additional details on the advance times relative to the two ML models are in Table 7).

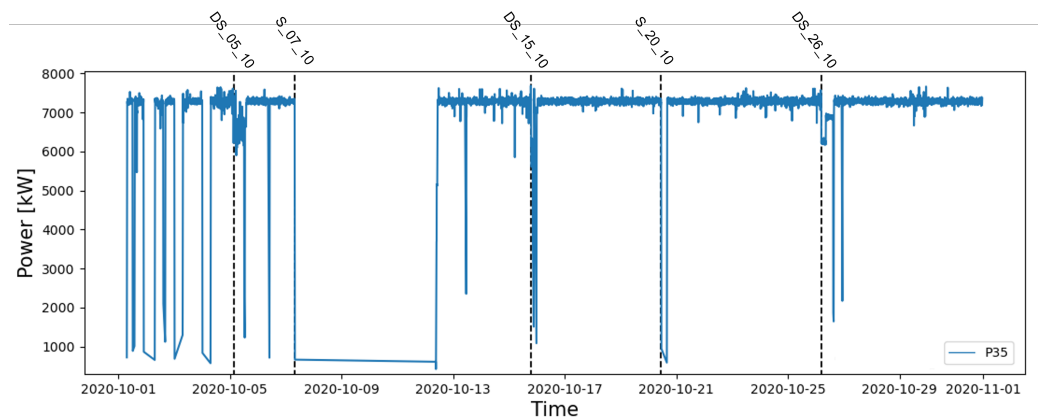


Figure 5. Active power in the reference period of October 2020 with details on the SCADA events recorded in that period (black dashed line).

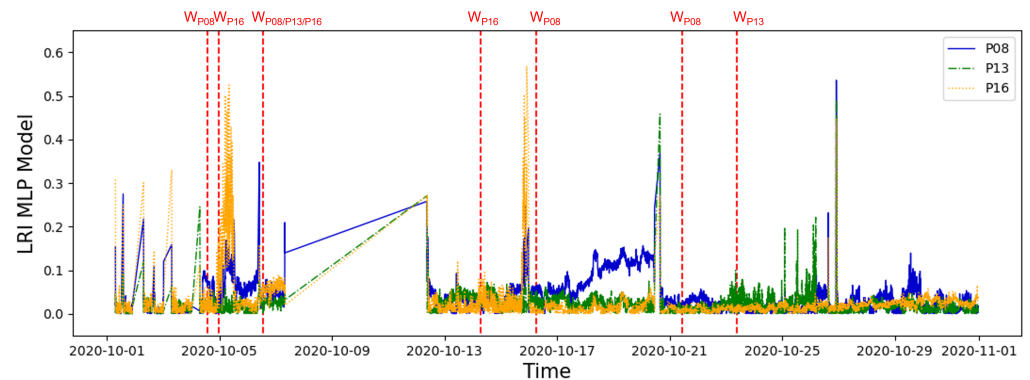


Figure 6. LRIs related to the parameters that caused a warning (red dashed line) after the application of the sliding threshold metric for the MLP model in October 2020.

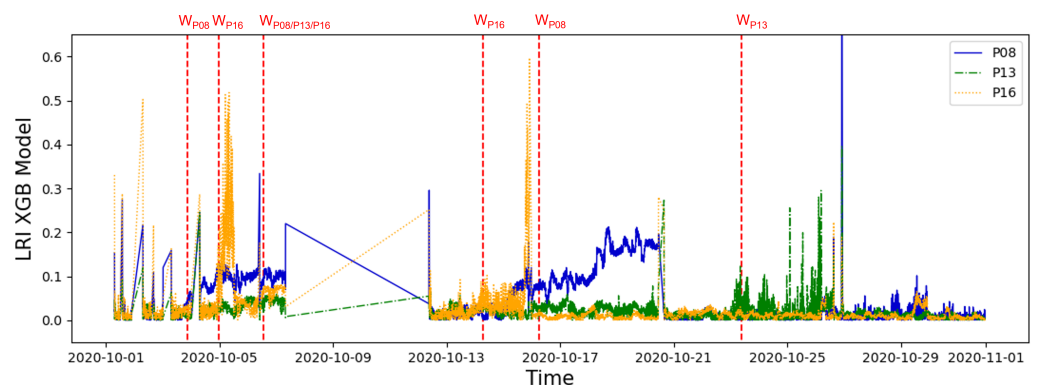


Figure 7. LRIs related to the parameters that caused a warning (red dashed line) after the application of the sliding threshold metric for the XGB model in October 2020.

Furthermore, while the warning on the S_15_10 event was triggered by the two models at the same time, the MLP model detected the anomaly related to event DS_05_10 about ten hours earlier than XGB. The third derating event followed by an engine shutdown was recorded on 26 October 2020 (event ID: DS_26_10) and concerned a high-temperature alarm on cylinder 5B (P13) detected on the same day. Furthermore, for this event, Figures 6 and 7 present a significant variation of the LRI for the parameter P13, constituting a specific precursor that results in a warning both for the MLP and XGB models on 23 October 2020, about three days in advance compared to the SCADA alarm.

Table 7. Comparison of detection performance of unscheduled downtime events (October–December 2020).

Event ID	XGB Results Detection (dd/mm/yy; hh/mm)	Anticipation (hh)	Precursors ID	MLP Results Detection (dd/mm/yy; hh/mm)	Anticipation (hh)	Precursors ID
DS_05_10	4 October 2020; 23:30	4	P16	4 October 2020; 13:45	14	P16
S_07_10	3 October 2020; 20:25	84	P08	4 October 2020; 17:20	62	P08
	6 October 2020; 13:15	18	P13, P16	6 October 2020; 14:05	17	P13, P16
DS_15_10	14 October 2020; 06:40	37	P16	14 October 2020; 08:10	34	P16
S_20_10	16 October 2020; 06:00	101	P08	16 October 2020; 07:05	100	P08
DS_26_10	23 October 2020; 09:30	67	P13	23 October 2020; 10:45	65	P13
S_13_11	10 November 2020; 12:55	69	P08	10 November 2020; 14:15	68	P08
S_19_11	14 November 2020; 00:25	123	P04, P08	14 November 2020; 01:05	122	P04, P08
DS_16_12	4 December 2020; 00:10	299	P28, P29, P30	4 December 2020; 01:25	298	P29, P31
	4 December 2020; 00:20	299	P04, P31, P32	-	-	-
S_21_12	16 December 2020; 12:05	114	P04, P08	16 December 2020; 13:00	113	P31, P32

The advances found before the emergency stops on 7 October 2020 (event ID: S_07_10) and 20 October 2020 (event ID: S_20_10) are of particular interest since they are not associated with a specific SCADA anomaly alarm on a component of the gas genset. In correspondence to these unscheduled downtimes, both ML models showed an anomaly on the LRI of cylinder temperature (P08), which caused a warning three days in advance of the first event (84 h for XGBoost and 62 h for MLP). Subsequently, the indicator of parameter P08 returned to normal values after the maintenance intervention, as visible in the active power plot in Figure 5), and then deviated again from 16 October 2020 (see Figures 6 and 7) until the emergency stop on 20 October 2020.

In a similar fashion, Figures 8–10 compare the results of the CBM method during November and December 2020, during which four significant unscheduled downtimes were reported by the SCADA system. Details on the event log can be found in Table 3.

Those events include three emergency stop alarms recorded, respectively, on 13 November 2020 (event ID: S_13_11), 19 November 2020 (event ID: S_19_11) and 21 December 2020 (event ID: S_21_12) as well as a shutdown transient due to an anomaly found on the generator temperature on 16 December 2020 (event ID: DS_16_12). From a global analysis of the LRI trends, shown in Figures 9 and 10, different anomalies were detected during the observed period in the engine cylinders and generator. In particular, previously found anomalies on the cylinder exhaust temperature, correlated with two long outages in October 2020, recurred from 13 November 2020, when a warning on the involved parameter was triggered by both MLP (Figure 9) and XGBoost (Figure 10). This significant deviation of the P08 parameter indicator persisted for about three days until an emergency stop was recorded on 13 November 2020.

Immediately after this 4-h engine outage, both models detected a new significant anomaly on P08, also involving other cylinders' temperatures and anticipating the failure detected by SCADA on 19 November 2020 (event ID: S_19_11). Of particular interest are the results related to the remaining two significant events recorded by the SCADA in December

2020, namely, DS_16_12 and S_21_12. In fact, the warnings detected so far by XGBoost and MLP were always triggered by the same parameters (with some differences only in the advance times with respect to the SCADA events), while in these two cases, different precursors emerged from the models.

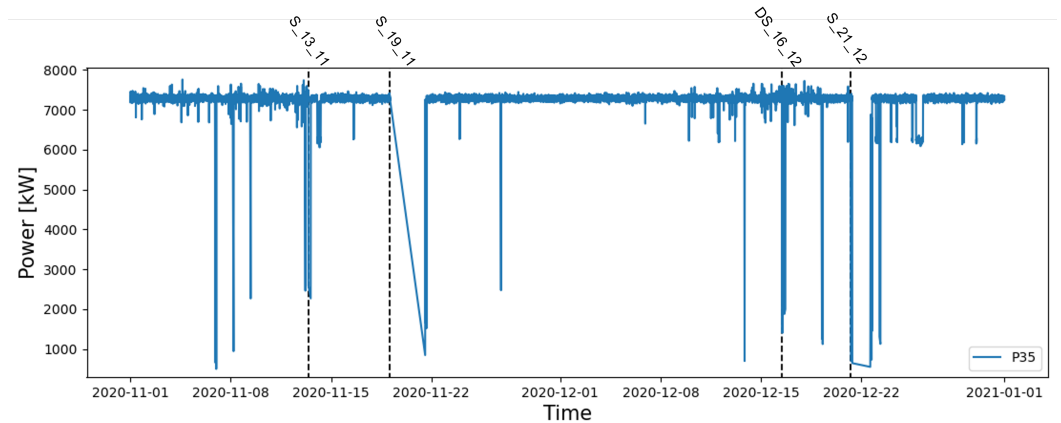


Figure 8. Active power for the period of November and December 2020, with details on the SCADA events recorded in that period (black dashed line).

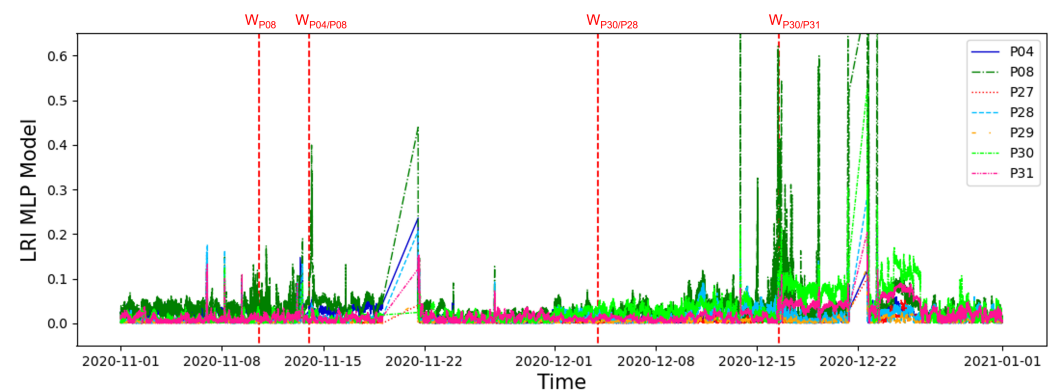


Figure 9. LRIs related to the parameters that generated a warning (red dashed line) after the application of the sliding threshold metric for the MLP model in November and December 2020.

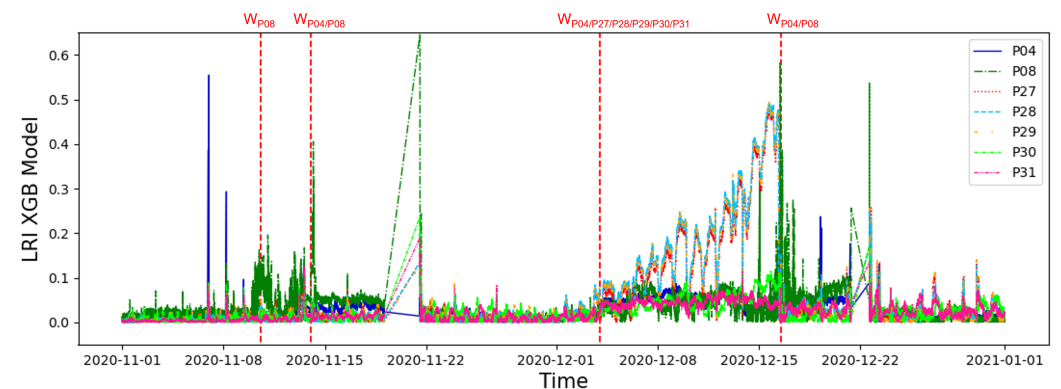


Figure 10. LRIs related to the parameters that generated a warning (red dashed line) after the application of the sliding threshold metric for the XGB model in November and December 2020.

In particular, XGBoost LRIs (Figure 9) highlighted, on 4 December 2020, a variation in the three temperatures of the generator-related variables, phases and bearings (P27–P31). This resulted in a warning that anticipates the SCADA log DS_16_12 by about twelve days. Comparing these results with those of the MLP model (Figure 10), the same

significant deviation was not noticed on the generator stator winding but only on the two generator bearings.

As for the unscheduled downtime of 21 December 2020, it was detected about 5 days in advance by both models, with different precursors: exhaust cylinder temperatures (P01-P19) for the XGB model and generator bearing temperatures (P27-P31) for the MLP model.

Finally, Table 7 summarizes the results discussed so far. In particular, the ability of each of the two ML models was assessed to identify specific precursors for major faults included in the SCADA log and then quantified the time of advance warning of the model relative to the occurrence of the reference SCADA alarm.

5. Conclusions

In this paper, an anomaly detection framework for the CBM of natural gas engines used in DH applications was presented. The framework exploited the use of signals collected by the SCADA system. The peculiarities of the framework reside in the PPS-inspired feature selection to reduce dataset dimensionality, the indifference to training dataset clustering to discriminate faults and normal operations and the management of time-series high-frequency information content directly filtering local residuals.

Two different models were tested to represent two different algorithm families: XGB in the symbolist family of decision trees and MLP in the connectivist family of neural networks. These models were trained to learn the regular behavior of the system based on a Leave-One-Out Cross-Validation approach and, based on the model reconstruction errors, a Local Residual Indicator (LRI) was defined for each monitored variable. Therefore, with the aim of triggering an early warning before the occurrence of faults, while limiting false alarms associated with instantaneous peaks in LRIs, a sliding threshold metric based on a moving average was adopted. In this way, a warning was triggered for the signals with the highest reconstruction error, to isolate the parameters mostly involved in the anomaly for troubleshooting purposes.

The proposed method was validated on 5 min SCADA data collected from a 7.5 MWe natural gas engine installed in the District Heating plant of Aosta city. The model was tested on anomalous periods selected using the SCADA event log. The results show that the proposed multivariate nowcasting approach allows the unveiling of hidden precursor dynamics that anticipate all the main fault events that occurred in the observed period. It is interesting to note that these anomalies were not detected by single-variable operational control approaches typical of SCADA systems.

In addition, even if both ML models anticipated the same faults with similar advance times, the better performance of XGB compared to MLP was evident in terms of the training customary scores for the nowcasting of single parameters (see Table 6). In particular, XGB paired with the two-stage threshold tuned with a persistence time of 6 h and time window size of 24 h provided fault anticipations ranging from 4 to 299 h. The framework proved to be fault agnostic because it detected ICE and generator anomalies.

In conclusion, the proposed solution presents a number of benefits due to its nature, which includes the ability to early detect anomalies in NG genset in DH networks, enabling the timely planning of corrective measures before major failures occur. This feature aligns with a CBM approach, where predictive maintenance strategies are adopted to ensure equipment performance and prevent unexpected downtime. Moreover, the proposed solution is cost-effective, as it works directly on the data sampled from the integrated SCADA systems. Unlike other systems that require additional intervention costs, the proposed solution operates directly on the available data and can be seamlessly integrated into the existing system.

The proposed solution employs a non-supervised approach that does not require labels to classify operational states during the training phase, which can be challenging to obtain. This feature makes the proposed solution highly versatile and adaptable to a wide range of systems and contexts. The methodological framework also introduces innovative solutions compared to the state of the art, including a feature selection phase

based on CPSS that optimizes the response times of the algorithm to obtain near real-time responses. Additionally, the training approach does not require a preliminary isolation of faulty conditions for the identification of the reference normal behavior model.

Finally, a post-processing of residuals is introduced through the use of a two-stage sliding threshold metric that provides nowcasting of alarms. Overall, the proposed solution offers a highly effective, efficient and cost-saving approach compared to the other systems and methods currently used in the industry. Future research could explore the potential of scaling up the solution for larger DH networks and testing its application in other domains.

Author Contributions: Conceptualization, A.C. and R.S.; methodology, V.F.B. and F.B.; formal analysis, V.F.B.; investigation, V.F.B., F.B. and F.A.T.; resources, R.S.; writing—original draft F.A.T. and F.B.; writing—review and editing, V.F.B. and F.B.; visualization, F.A.T.; supervision, A.C.; project administration, A.C. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a contract between Engie Servizi S.p.A. and DIMA, contract number 9149/20. The APC was funded by DIMA.

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

References

- Verda, V.; Colella, F. Primary energy savings through thermal storage in district heating networks. *Energy* **2011**, *36*, 4278–4286. [CrossRef]
- Dominković, D.F.; Stunjek, G.; Blanco, I.; Madsen, H.; Krajačić, G. Technical, economic and environmental optimization of district heating expansion in an urban agglomeration. *Energy* **2020**, *197*, 117243. [CrossRef]
- Cai, H.; Ziras, C.; You, S.; Li, R.; Honoré, K.; Bindner, H.W. Demand side management in urban district heating networks. *Appl. Energy* **2018**, *230*, 506–518. [CrossRef]
- Alexandrov, G.; Ginzburg, A. Anthropogenic impact of Moscow district heating system on urban environment. *Energy Procedia* **2018**, *149*, 161–169. [CrossRef]
- Milić, V.; Amiri, S.; Moshfegh, B. A systematic approach to predict the economic and environmental effects of the cost-optimal energy renovation of a historic building district on the district heating system. *Energies* **2020**, *13*, 276. [CrossRef]
- Newell, R.; Raimi, D.; Villanueva, S.; Prest, B. Global energy outlook 2021: Pathways from Paris. *Resour. Future Rep.* **2021**, *8*, 11–21.
- Bouckaert, S.; Pales, A.F.; McGlade, C.; Remme, U.; Wanner, B.; Varro, L.; D’Ambrosio, D.; Spencer, T. *Net Zero by 2050: A Roadmap for the Global Energy Sector*; Report; IEA: Paris, France, 2021. Available online: <https://www.iea.org/reports/net-zero-by-2050> (accessed on 13 March 2023).
- Sayegh, M.A.; Jadwiszczak, P.; Axcell, B.; Niemierka, E.; Bryś, K.; Jouhara, H. Heat pump placement, connection and operational modes in European district heating. *Energy Build.* **2018**, *166*, 122–144. [CrossRef]
- Levihn, F. CHP and heat pumps to balance renewable power production: Lessons from the district heating network in Stockholm. *Energy* **2017**, *137*, 670–678. [CrossRef]
- Wang, H.; Yin, W.; Abdollahi, E.; Lahdelma, R.; Jiao, W. Modelling and optimization of CHP based district heating system with renewable energy production and energy storage. *Appl. Energy* **2015**, *159*, 401–421. [CrossRef]
- IEA. *District Heating*; Report; IEA: Paris, France, 2021.
- Mendonça, P.; Bonaldi, E.; de Oliveira, L.; Lambert-Torres, G.; da Silva, J.B.; da Silva, L.B.; Salomon, C.; Santana, W.; Shinohara, A. Detection and modelling of incipient failures in internal combustion engine driven generators using electrical signature analysis. *Electr. Power Syst. Res.* **2017**, *149*, 30–45. [CrossRef]
- Yun, Q.; Zhang, C.; Ma, T. Fault diagnosis of diesel generator set based on deep believe network. In Proceedings of the second International Conference on Artificial Intelligence and Pattern Recognition, Beijing, China, 16–18 August 2019; pp. 186–190.
- Assuncao, F.d.O.; Borges-da Silva, L.E.; Villa-Nova, H.F.; Bonaldi, E.L.; Oliveira, L.E.L.; Lambert-Torres, G.; Teixeira, C.E.; Sant’Ana, W.C.; Lacerda, J.; da Silva Junior, J.L.M.; et al. Reduced Scale Laboratory for Training and Research in Condition-Based Maintenance Strategies for Combustion Engine Power Plants and a Novel Method for Monitoring of Inlet and Exhaust Valves. *Energies* **2021**, *14*, 6298. [CrossRef]
- Basurko, O.C.; Uriondo, Z. Condition-Based Maintenance for medium speed diesel engines used in vessels in operation. *Appl. Therm. Eng.* **2015**, *80*, 404–412. [CrossRef]
- Vera-García, F.; Pagán Rubio, J.A.; Hernández Grau, J.; Albaladejo Hernández, D. Improvements of a failure database for marine diesel engines using the RCM and simulations. *Energies* **2019**, *13*, 104. [CrossRef]
- Aliramezani, M.; Koch, C.R.; Shahbakhti, M. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: A review and future directions. *Prog. Energy Combust. Sci.* **2022**, *88*, 100967. [CrossRef]
- Ntakolia, C.; Anagnostis, A.; Moustakidis, S.; Karcianas, N. Machine learning applied on the district heating and cooling sector: A review. *Energy Syst.* **2021**, *13*, 1–30. [CrossRef]

19. Mbiydzennyuy, G.; Nowaczyk, S.; Knutsson, H.; Vanhoudt, D.; Brage, J.; Calikus, E. Opportunities for machine learning in district heating. *Appl. Sci.* **2021**, *11*, 6112. [[CrossRef](#)]
20. Baranowski, J.; Bania, P.; Prasad, I.; Cong, T. Bayesian fault detection and isolation using Field Kalman Filter. *EURASIP J. Adv. Signal Process.* **2017**, *2017*, 1–11. [[CrossRef](#)]
21. Flett, J.; Bone, G.M. Fault detection and diagnosis of diesel engine valve trains. *Mech. Syst. Signal Process.* **2016**, *72*, 316–327. [[CrossRef](#)]
22. Jung, D. Data-driven open-set fault classification of residual data using Bayesian filtering. *IEEE Trans. Control. Syst. Technol.* **2020**, *28*, 2045–2052. [[CrossRef](#)]
23. Czech, P.; Mikulski, J. Application of Bayes classifier and entropy of vibration signals to diagnose damage of head gasket in internal combustion engine of a car. In Proceedings of the International Conference on Transport Systems Telematics, Katowice/Krakow/Ustron, Poland, 20–25 October 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 225–232.
24. Zhang, F.; Jiang, M.; Zhang, L.; Ji, S.; Sui, Q.; Su, C.; Lv, S. Internal combustion engine fault identification based on FBG vibration sensor and support vector machines algorithm. *Math. Probl. Eng.* **2019**, *2019*, 8469868. [[CrossRef](#)]
25. Dandare, S.; Dudul, S. Support vector machine based multiple fault detection in an automobile engine using sound signal. *J. Electron. Electr. Eng.* **2012**, *3*, 59–63.
26. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review. *IET Renew. Power Gener.* **2017**, *11*, 382–394. [[CrossRef](#)]
27. Helbing, G.; Ritter, M. Deep Learning for fault detection in wind turbines. *Renew. Sustain. Energy Rev.* **2018**, *98*, 189–198. [[CrossRef](#)]
28. Liu, Y.; Chang, W.; Zhang, S.; Zhou, S. Fault diagnosis and prediction method for valve clearance of diesel engine based on linear regression. In Proceedings of the 2020 Annual Reliability and Maintainability Symposium (RAMS), Palm Springs, CA, USA, 27–30 January 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
29. Bryg, D.J.; Mink, G.; Jaw, L.C. Combining lead functions and logistic regression for predicting failures on an aircraft engine. In Proceedings of the Turbo Expo: Power for Land, Sea, and Air, Berlin, Germany, 9–13 June 2008; Volume 43123, pp. 19–26.
30. Singh, D.; Kumar, M.; Arya, K.; Kumar, S. Aircraft Engine Reliability Analysis using Machine Learning Algorithms. In Proceedings of the 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 26–28 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 443–448.
31. Maraini, D.; Simpson, M.; Brown, R.; Poporad, M. Development of a Data-driven Model for Marine Gas Turbine (MGT) Engine Health Monitoring. In Proceedings of the Annual Conference of the Prognostics and Health Management Society, Paris, France, 2–5 May 2018.
32. Chen, M.; Li, Z.; Lei, X.; Liang, S.; Zhao, S.; Su, Y. Unsupervised Fault Detection Driven by Multivariate Time Series for Aeroengines. *J. Aerosp. Eng.* **2023**, *36*, 04022129. [[CrossRef](#)]
33. Deon, B.; Cotta, K.; Silva, R.; Batista, C.; Justino, G.; Freitas, G.; Cordeiro, A.; Barbosa, A.; Loução Jr, F.; Simioni, T.; et al. Digital twin and machine learning for decision support in thermal power plant with combustion engines. *Knowl.-Based Syst.* **2022**, *253*, 109578. [[CrossRef](#)]
34. Braei, M.; Wagner, S. Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv* **2020**, arXiv:2004.00433.
35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
36. Kolmogorov, A. On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Inf. Theory* **1956**, *2*, 102–108. [[CrossRef](#)]
37. Wetschoreck, F.; Krael, T.; Krishnamurthy, S. Online Repository. 2021. Available online: <https://github.com/8080labs/ppscore/releases> (accessed on 13 March 2023).
38. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
39. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.
40. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
41. Miele, E.S.; Bonacina, F.; Corsini, A. Deep anomaly detection in horizontal axis wind turbines using Graph Convolutional Autoencoders for Multivariate Time series. *Energy AI* **2022**, *8*, 100145. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.