applied sciences

MDPI

Article

Data-Quality Assessment for Digital Twins Targeting Multi-Component Degradation in Industrial Internet of Things (IIoT)-Enabled Smart Infrastructure Systems

Atuahene Kwasi Barimah [1,*], Octavian Niculita [1], Don McGlinchey [2] and Andrew Cowell [2]

[1] Department of Applied Science, School of Computing, Engineering and Built Environment, Glasgow Caledonian University, Glasgow G4 0BA, UK; octavian.niculita@gcu.ac.uk
[2] Department of Mechanical Engineering, School of Computing, Engineering and Built Environment, Glasgow Caledonian University, Glasgow G4 0BA, UK; d.mcglinchey@gcu.ac.uk (D.M.); a.cowell@gcu.ac.uk (A.C.)
* Correspondence: abarim300@caledonian.ac.uk; Tel.: +44-141-273-1816

Featured Application: The paper highlights the steps taken in checking the required data quality (both real and synthetic) before it is used for the development of services in the context of IIoT-enabled smart infrastructure systems. A case study of a scaled-down version of a water distribution system will be presented in detail and data from healthy and faulty conditions will be used to demonstrate the details of the data qualification process and the impact on various health assessment techniques meant to support fault detection and isolation of single and multi-component degradation scenarios. The paper also proposes an IIoT architecture for the instantiation of measurement system analysis.

Abstract: In the development of analytics for PHM applications, a lot of emphasis has been placed on data transformation for optimal model development without enough consideration for the repeatability of the measurement systems producing the data. This paper explores the relationship between data quality, defined as the measurement system analysis (MSA) process, and the performance of fault detection and isolation (FDI) algorithms within smart infrastructure systems. This research employs a comprehensive methodology, starting with an MSA process for data-quality evaluation and leading to the development and evaluation of fault detection and isolation (FDI) algorithms. During the MSA phase, the repeatability of a water distribution system's measurement system is examined to characterise variations within the system. A data-quality process is defined to gauge data quality. Synthetic data are introduced with varying data-quality levels to investigate their impact on FDI algorithm development. Key findings reveal the complex relationship between data quality and FDI algorithm performance. Synthetic data, even with lower quality, can improve the performance of statistical process control (SPC) models, whereas data-driven approaches benefit from high-quality datasets. The study underscores the importance of customising FDI algorithms based on data quality. A framework for instantiating the MSA process for IIoT applications is also suggested. By bridging data-quality assessment with data-driven FDI, this research contributes to the design of digital twins for IIoT-enabled smart infrastructure systems. Further research on the practical implementation of the MSA process for edge analytics for PHM applications will be considered as part of our future research.

Keywords: digital twins; industrial internet of things (IIoT); instrumentation; data quality; statistical process control; machine learning

Citation: Barimah, A.K.; Niculita, O.; McGlinchey, D.; Cowell, A. Data-Quality Assessment for Digital Twins Targeting Multi-Component Degradation in Industrial Internet of Things (IIoT)-Enabled Smart Infrastructure Systems. *Appl. Sci.* **2023**, *13*, 13076. https://doi.org/10.3390/app132413076

Academic Editor: Dimitris Mourtzis

Received: 30 October 2023
Revised: 1 December 2023
Accepted: 4 December 2023
Published: 7 December 2023

1. Introduction

To meet the evolving needs of a dynamic, global urban space, infrastructure assets supporting modern cities should make use of technologies that drive the operations and

predictive maintenance activities enabled by the capabilities of the industrial internet of things (IIoT). In recent years, more and more articles have discussed the concepts of digital twins and their use in the context of smart cities. The sense–acquire–transform–analyse–act sequence is not new, and it underpinned the topic of condition monitoring for the past two decades. The adoption of the related concepts of condition monitoring (CM), industrial internet of things (IIoT), digital twins (DTs), and cyber–physical systems (CPS) in modern city infrastructure applications is often described as the fourth generation of the industrial revolution. The central theme of this so-called revolution is the integration of the physical with the digital world and the visualisation of insights and realisation service offerings that can be obtained therefrom. Condition monitoring is the process of monitoring a condition parameter in machinery to identify a significant change, which is indicative of degradation/anomaly.

The IIoT can be defined as the connection of industrial equipment to networked computing resources to collect data and use these data to drive relevant outcomes/services. In a similar vein, a digital twin can be described as a virtual data model that represents an actual physical entity in digital space—within the confines of a defined use case or experimental frame. Cyber–physical systems can comparably be described as systems that integrate the dynamics of physical processes with computation, networking, and decision-making capabilities. These four concepts are so closely related that they are sometimes confused with each other and used interchangeably, although, over the years, academia and industry have tried to define processes for their design and implementation, as well as standards and recommended practices for their realisation and deployment. Fuller et al. [1] provide a tabulated review of recent research works relating to digital twins and highlight some common misconceptions with other cognate notions of Industry 4.0. Tao et al. [2] discuss the correlation and comparisons between DTs and CPS, whereas Lu et al. [3] distinguish between all four concepts of CM, IIoT, DTs, and CPS and explain the interactions between them.

As stated in [4–6], the IIoT provides the data framework upon which most industrial digital twins are built. Thus, a thorough understanding of the IIoT application development process is crucial in implementing digital twins. In the built environment space, the concept of building information modelling (BIM) has begun to be discussed more often in recent years in relation to DTs and CPS [7]. If BIM is typically employed as designed/as built during the delivery stages of a project supporting commissioning activities and facilities management, DT capabilities tap into the asset management, predictive maintenance, what-if analysis, and a diverse range of simulations supporting the optimisation of operations. At the same time, a different perspective on DT-related research actively discussed in various working groups is attributed to the definitions of layers describing the digital counterpart of a physical asset. These layers refer to data services (DS), integration (I1), intelligence (I2), user experience (UX), management (M), and trustworthiness (T) [8]. Datasets supporting the development of DTs for smart infrastructure systems exist, and some can certainly support the instantiation of the DT layers mentioned above. An example of such a dataset is the one generated by [9] to target leak detection and localisation in water distribution systems.

Also, several authors have explored different fault diagnosis and prognostics approaches for various systems with the use of machine learning as a more recent phenomenon. Fernandes et al. [10] provide a systematic review of the application of machine learning techniques in real industrial use cases for fault diagnosis and prognosis. The research highlighted in their work shows that most supervised model development approaches rely on datasets from a measurement system. This is not uncommon in the development and application of data-driven methods for diagnostic or prognostic services, as seen in [11–13]. The approach used for data preparation in the development of analytics for prognostic and health management (PHM) applications depends on the nature of the training dataset used. Where data preparation methods focus on the distribution of the training dataset, data pre-processing paves the way for feature extraction, feature selection, and model training [14,15]. However, this approach does not consider the performance

of an asset's measurement system or its impact on the development of the analytics for PHM applications. This paper defines a measurement system analysis instantiation process for developing fault diagnosis and isolation (FDI) algorithms to detect multi-component degradation in a water distribution system. The main contributions of this paper are summarised below:

1. Definition of a measurement system analysis process for developing fault detection and isolation algorithms for multi-component degradation;
2. Investigation of the impact of data quality on the development of fault detection and isolation algorithms for multi-component degradation for IIoT-enabled smart infrastructure systems;
3. Delineation of an application development process with clear steps for the quantification of data quality before it is used for the design and development of services for IIoT-enabled water distribution systems;
4. Demonstration of a practical implementation of the application of a measurement system assessment and offering of perspectives on the impact of data quality on SPC applications, an ensemble model of ML techniques, and neural networks, which can act as a reference for data analysts operating in the DT/IIoT space.

2. Methodology

2.1. Methodology for the Assessment of Data Quality

2.1.1. Measurement System Analysis (MSA)—Methodology Overview

Measurement system analysis (MSA) is generally a set of procedures used to determine the variation caused by a measurement system and whether the measurement data are valid before being used for the construction of services delivered by a DT. MSA aims to separate the variation among devices being measured from the error in the measurement system [16]. Understanding and minimising the sources of variation in the measurement process allows organisations to make more informed decisions and improve the quality of their products or services. An MSA process was defined to determine the level of repeatability by a water distribution testbed's measurement system to aid in identifying the impact of the data quality on analytics as a service. The analysis starts with defining a process for capturing healthy and faulty condition data over one month by different operators. A relative pooled variance, which computes a combined measure of variability by considering both the sample sizes and the variances of the groups, was used to determine whether there was variability in the healthy and faulty condition data over the period in which data were recorded. To determine the source of data variation when more than 10% relative pooled standard deviation was observed, a gage repeatability and reproducibility (gage R&R) approach was adopted to determine the variation in the process data associated with the process and the measurement system. The earth mover's distance (EMD) metric was also used to complement the gage R&R analysis by determining the similarity between datasets by measuring the distance between the probability distributions of the measured process data. A data-quality (DQ) score between 0 and 1 was then defined as an average of the scaled outcomes of both the gage R&R analysis (repeatability) and the EMD score (similarity metric), where a DQ score of 1 represents a highly repeatable measurement system.

2.1.2. Description of the System under Investigation

The investigation was conducted on a testbed (see Figure 1) capable of generating data meant to support the analysis of the dynamic behaviour of a water distribution system undergoing multi-component degradation. This hydraulic system comprises critical components, including a main supply tank, an external gear pump, and an induction motor responsible for driving the pump. The rotational speed of both the motor and the pump is regulated by a variable speed drive (VSD). The system also features a solenoid shut-off valve (SHV) and five direct proportional valves (DPV1 to DPV5) added to the system to support, in a controlled manner, the degradation phenomena affecting five distinct

components. Data collection is facilitated by pressure transmitters (P1, P2, P3, P4, and P5), turbine flow meters (f1 and f2), and a laser sensor to measure the pump's speed.

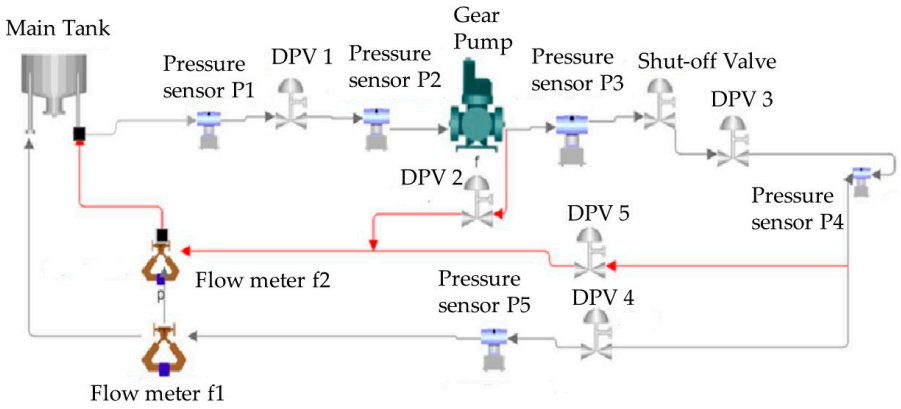

Figure 1. Water distribution system testbed schematic.

System components are connected with PVC tubing, and a finger valve is used for tank isolation when needed. In the context of fault simulation, specific control valves were manipulated to emulate fault conditions. For instance, DPV1 represented a clogged suction filter, fully open at 0% fault severity, whereas DPV2 simulated pump discharge side leakage and was fully closed at 0% fault severity. The SHV solenoid valve remained open, and DPV3, emulating a blocked or degraded shut-off valve, was fully open at 0% fault severity. DPV4 represented a clogged nozzle, also fully open at 0% fault severity, whereas DPV5, simulating downstream pipe leakage, was fully closed with 0% fault severity. For clarity, Table 1 summarises the default operational states as well as the fault emulation mechanism of the system's control valves and also provides their corresponding fault codes.

Table 1. Healthy condition operating state of the system's control valves and associated fault codes.

Component	Fault Emulation	Healthy State/Fault Emulation Mechanism	Fault Code
Filter	DPV 1	Fully open/gradually closing	FC1
Pump	DPV 2	Fully closed/gradually opening	FC2
Valve	DPV 3	Fully open/gradually closing	FC3
Nozzle	DPV 4	Fully open/gradually closing	FC4
Pipe	DPV 5	Fully closed/gradually opening	FC5

2.1.3. Process Data Capture

Data for the fuel system process was recorded within four weeks for healthy condition (HC), single-component degradation (SCD), and multi-component degradation (MCD) scenarios between pump speeds of 700 rpm and 950 rpm in intervals of 50 rpm. The SCD process data represent the degradation of individual components (see Table 2) with pressure and flow measurements at $P_1, P_2 \ldots . P_5$ and f_1, f_2, respectively. Data logging for each faulty condition scenario started at least 10 min after the process had reached steady-state conditions or when there was a step change in pump speed or a change in the failure condition scenario, with each data file having different sample sizes. Figures 2 and 3 show the distribution of process parameters for a filter-clogging scenario with a gradual closure of DPV1 (see Figure 1) and a water distribution system under the healthy condition, respectively.

Table 2. Summary of data capture process.

Test Period	4 Consecutive Weeks
Faulty condition scenarios (total No. of tests)	FC0—healthy condition (24) FC1—clogged filter (24) FC2—degraded pump (24) FC3—blocked valve (24) FC4—blocked nozzle (24) FC5—leaking pipe (24)
Pump speed (rpm)	700/750/800/850/900/950

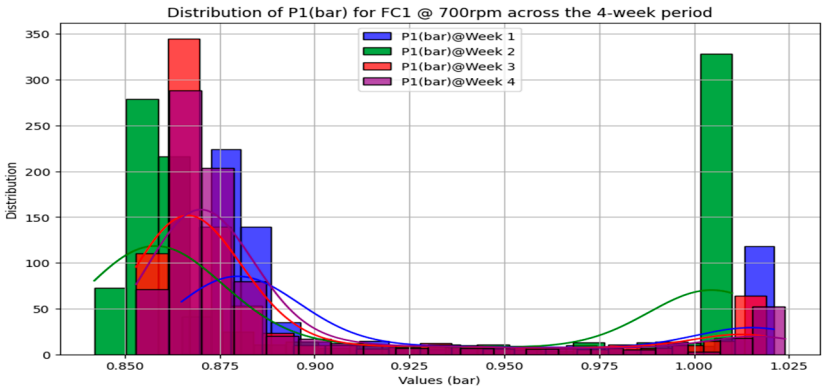

Figure 2. Distribution of data from pressure sensor 1 for a filter-clogging scenario with a pump at 700 rpm.

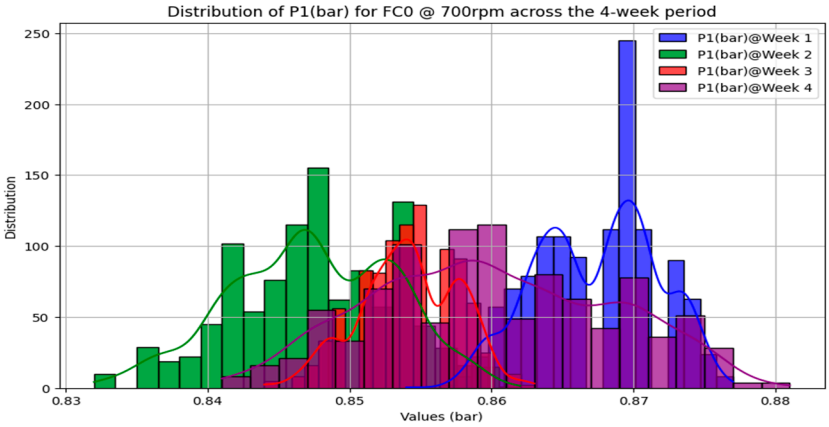

Figure 3. Distribution of data from pressure sensor 1 for a healthy condition scenario with a pump at 700 rpm.

2.1.4. Synthetic Data Generation

Synthetic data with varying levels of repeatability were generated from the distributions of the data captured over the 4 weeks. The rationale for generating synthetic data with varying levels of repeatability was to determine the impact of measurement system repeatability on the development of fault detection and isolation algorithms for multi-component degradation scenarios. To achieve this, a conditional tabular generative adversarial network (CTGAN) [17] was used to generate synthetic data from data captured in Table 2 with random hyperparameters in order to increase the variation in the recorded data at each sensor measurement. This enabled the generation of different datasets related to the same process dynamics for the training of fault detection and isolation algorithms.

2.1.5. Relative Pooled Variance

To calculate the relative pooled variance [18], the variance S^2 of each group of data in the dataset was multiplied by its degrees of freedom $(n_i - 1)$. These values were then added together for all groups. To obtain the relative pooled variance, the sum was divided by the total degrees of freedom (the sum of the individual degrees of freedom for each group). Equations (2) and (3) were used to determine the pooled standard deviation and relative pooled standard deviation, respectively.

$$Standard\ deviation = \sqrt{Variance} \tag{1}$$

Equation (1): standard deviation.

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2 + \ldots + (n_k - 1){S_k}^2}{n_1 + n_2 + .. + n_k - k}} \tag{2}$$

Equation (2): pooled standard deviation.

$$RPSD = \frac{SD_{pooled}}{mean} \tag{3}$$

Equation (3): relative pooled standard deviation.

2.1.6. Gage R&R Analysis

In this work, the gage R&R [19] approach was adopted to determine the variation in the process data associated with the process dynamics and the measurement system. The faulty conditions at different pump speeds were treated as parts in a gage R&R analysis, with the weekly data recording as replications and the test campaign as the operator (see Table 3). The grand mean for the process parameters of each faulty condition scenario (see Equation (4)) and the total sum of squares were determined to gauge the overall data variability. Subsequently, the operator-related variance was determined as well as the degrees of freedom, which in this case was the number of operators. Repeatability as a function of variation in the data was then determined (see Equation (5)) and scaled between 0 and 1.

$$FCX^{week}{}_{pump\_speed} = \sum mean(P1, P2, P3, P4, P5, F1, F2) \tag{4}$$

Equation (4): process data mean.

Table 3. Data for gage R&R analysis.

Week	Nomenclature Test	FC0_700	.....	FC5_950
1	Test campaign 1	$FC0^1{}_{700}$	.....	$FC5^1{}_{950}$
2	Test campaign 2	$FC0^2{}_{700}$	.....	$FC5^2{}_{950}$
3	Test campaign 3	$FC0^3{}_{700}$	.....	$FC5^3{}_{950}$
4	Test campaign 4	$FC0^4{}_{700}$	.....	$FC5^4{}_{950}$

$$Repeatability = 1 - \left(Mean\ sum\ of\ squares\ (operator) \Big/ Total\ sum\ of\ squares \right) \tag{5}$$

Equation (5): repeatability score.

2.1.7. Earth Mover's Distance

The earth mover's distance (EMD) is used to measure the discrepancy between two probability distributions and provides a way to quantify the minimum cost of transforming one distribution into another [20]. In this paper, the distributions of sensor data measured at $P_1, P_2 ..... P_5$ and f_1, f_2 were compared and the Wasserstein distance was

computed for each sensor measurement across the 4 weeks for each faulty condition. Equation (6) was used to compute a similarity metric scaled between 0 and 1, using the average EMD for faulty condition scenarios FC0, FC1, FC2, FC3, FC4, and FC5. A value of 1 represents a significant similarity between sensor data across the weeks and 0 represents dissimilar datasets across the week by the same sensor.

$$Similarity\ metric = 1 - \frac{Average\ EMD}{1 + Average\ EMD} \quad (6)$$

Equation (6): similarity metric for process data.

2.2. Component Degradation Characterisation

The nature of filter degradation, i.e., clogging over time due to the accumulation of particles or debris [21] (FC1), was emulated on the testbed by using the position of a valve (DPV1) that was proportional to the level of severity. In the case of the pump (FC2), pump failures can result from motor or drive issues, such as electrical faults, mechanical wear, or overheating, or leaks leading to a complete breakdown of the pumping system [22]. A degraded pump with leaks was simulated on the testbed with a change in the opening of a bypass valve (DPV2) connected to the main line (see Figure 1). Sediment build-up [23] was also emulated as the degradation of the valve (FC3) by a gradual closure in DPV3 or a step change in the valve opening. In the case of the nozzle (FC4), represented by DPV4, partial blockage generated by solid particle contaminants present in the fluid [24] was emulated by a gradual closure of the valve. Finally, a leaking pipe [15] faulty condition scenario (FC5) was emulated via a step change in the DPV5 valve opening, which offered the possibility of creating fine leaks or a significant pipe burst. In this paper, only two faulty conditions occurring simultaneously for the MCD case were considered, with each valve associated with each component under consideration set at a specific valve opening.

2.3. Fault Detection and Isolation (FDI)

2.3.1. Test Degradation Scenarios

Benchmarking was carried out using a series of test datasets recorded to assess the performance of FDI algorithms in the context of multi-component degradation. Table 4 below describes the nature of the test data with the components under consideration for the multi-component degradation scenarios at different pump speeds as well as their levels of severity. For example, T14 captured data from a scenario where the system (running at 950 rpm) was affected by a combination of a degraded pump and nozzle; the pump's level of severity was emulated by setting the DPV2 at 45% open, whereas the degradation of the nozzle was emulated by swiping the entire opening envelope of the DPV4 (100% open highlighting a healthy nozzle and 30% opening emulating a high-severity level of degradation). Also, Appendix A.1 (Figure A1) shows the distribution of the MCD test degradation scenarios in Table 4 used to test the performance of the FDI algorithms. The levels of severity were categorised into five groups, with below 20% defined as healthy, between 20% and 40% defined as low severity, between 40% and 60% defined as medium severity, between 60% and 80% defined as high severity, and above 80% defined as high failure risk. The FDI algorithms were expected to classify or predict any instances above 20% severity as faulty in both multi-component and single-component scenarios. To better clarify how the deviation from normality occurred, Figures 4 and 5 show how the distribution of the process parameters of the water distribution system deviated in test degradation scenarios T13 and T17, respectively.

Table 4. Test degradation scenarios (the fault in component 1 is initiated before that of component 2).

MCD Test No. (Total No. of Tests)	Degradation Level (Component 1) (in Percentage of DPV Opening)	Degradation Level (Component 2) (in Percentage of DPV Opening)	Operational Speed Range (rpm)	Fault Combination
T13 (1)	Pump (medium severity) at 45%	Constant degradation of nozzle (30–100% severity)	700	FC2 and FC4
T14 (1)	Pump (medium severity) at 45%	Constant degradation of nozzle (30–100% severity)	950	FC2 and FC4
T15 (1)	Filter (high severity) at 68%	Healthy condition (0% severity)	700 to 950	FC1 only
T16 (1)	Pump (medium severity) at 50%	Nozzle (medium severity) at 60%	700 to 950	FC2 and FC4
T17 (1)	Constant degradation of pump (0–100% severity)	Constant degradation of pipe (0–100% severity)	800	FC2 and FC5
T18 (1)	Intermittent faults for the pump between a 45% and 60% level of severity	Healthy condition (0% severity)	850	FC2 only
T19 (1)	Constant degradation of pump (0–100% severity)	Constant degradation of valve (30–100% severity)	850	FC2 and FC3
T20 (1)	Pump (medium severity) at 55%	Nozzle (high severity) at 70%	700 to 950	FC2 and FC4

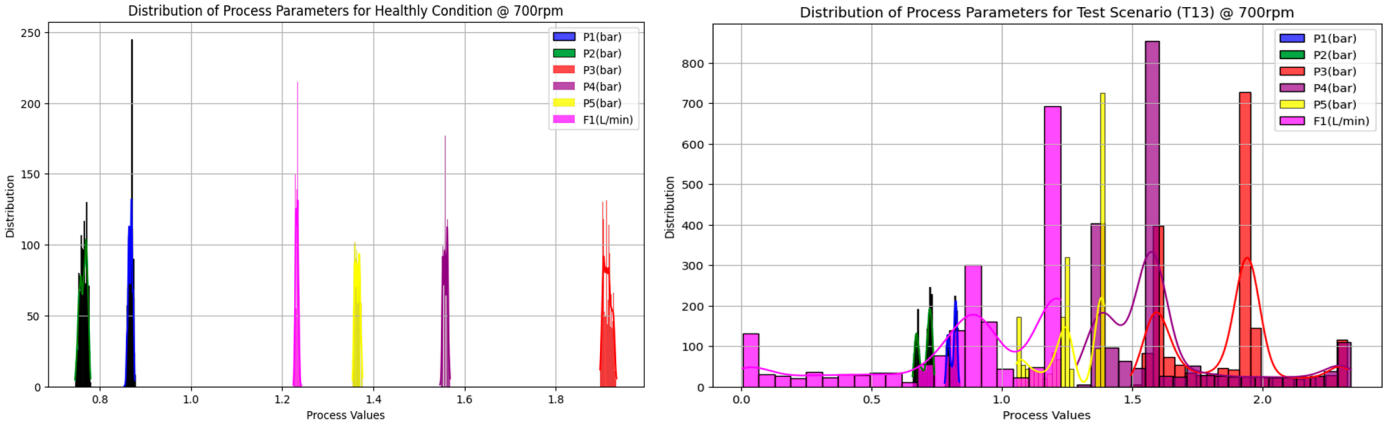

Figure 4. Deviation of process parameters from healthy condition for T13.

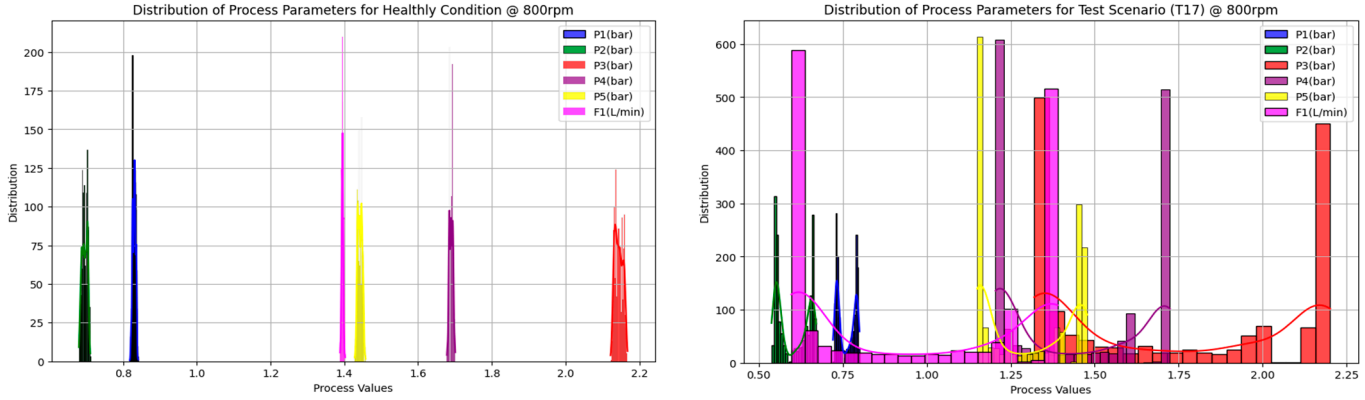

Figure 5. Deviation of process parameters from healthy condition for T17.

2.3.2. Statistical Process Control (SPC) Approach

Statistical process control (SPC) is a quality-control method used to monitor and control processes to ensure they operate consistently within specified limits. It involves the

use of statistical techniques to analyse process data, identify variations, and take corrective actions when necessary. Threshold testing is a critical aspect of statistical process control (SPC) used to determine whether a process is operating within acceptable limits. In SPC, control charts are employed to monitor process data over time, and threshold testing involves setting limits or thresholds on these charts to distinguish between common cause variation and special cause variation [25]. In this report, Equation (7) was used to calculate the upper and lower bounds for pressure and flow measurements at each pump speed for each faulty condition scenario, where μ and σ are the mean and standard deviation of the sensor measurements, respectively. Equation (8) was also used to develop the SPC table at each pump speed for each faulty condition scenario, which was then used to set the thresholds for each component failure. T_1 and T_2 represent the lower and upper bounds, respectively, of the process dynamic faulty condition scenarios FC1, FC2, FC3, FC4, and FC5, and the logic condition returned 1 when true and 0 when false using Equation (9). The logic condition of each faulty condition at time t_1 returned a response of "true" when $T_1 \leq x_1 \leq T_2$, where x_1 is the process parameter at t_1. An SPC surface was then developed for every faulty condition scenario at pump speeds of between 700 and 950 rpm at intervals of 50 rpm and stored in an SPC model repository for FDI predictions. For instance, the SPC model repository contained an SPC surface such as the one shown in Figure 6 for sensor measurements at P3 for the healthy condition scenario of the system. It was then used to define the distribution of the data at that sensor measurement for a particular component condition.

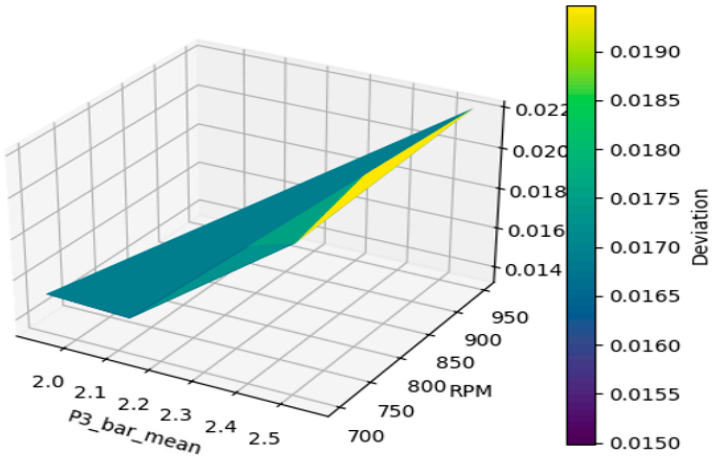

Figure 6. SPC surface for sensor data (historical) at P3 for the healthy condition.

$$\begin{aligned} Upper_{bound}(UB) &= \mu + 3\sigma \\ Lower_{bound}(LB) &= \mu - 3\sigma \end{aligned} \tag{7}$$

Equation (7): lower and upper bounds.

$$SPC\_table\_FCX = f_{UB,LB}{}^{pump\_speed}(P1, P2, P3, P4, P5, f1) \tag{8}$$

Equation (8): threshold testing table for process parameters for each process condition.

$$\left\{ \begin{array}{c} Pump\_Speed \\ AND \\ SPC\_table\_FCX_{LB_{P1}} \leq Test_{P1} \leq SPC\_table\_FCX_{UB_{P1}} \\ . \\ . \\ SPC\_table\_FCX_{LB_{P5}} \leq Test_{P5} \leq SPC\_table\_FCX_{UB_{P5}} \\ AND \\ SPC\_table\_FCX_{LB_{f1}} \leq Test_{F1} \leq SPC\_table\_FCX_{UB_{f1}} \end{array} \right\} \tag{9}$$

Equation (9): SPC algorithm for fault detection and isolation (FDI).

2.3.3. Data-Driven Approach

The data-driven approach for fault detection and isolation proposed in this paper was two-fold. The first part was an ensemble of classification algorithms [26] for detecting faulty condition scenarios. A range of classification models, including logistic regression [27], decision tree [28], random forest [29], Gaussian naive Bayes [30], K-nearest neighbours [31], support vector machine [32], gradient boosting [33], and AdaBoost [34], were instantiated. These classifiers were fitted to training data and evaluated based on their performance for each of the key components (see Table 1) of the system on test degradation data in terms of prediction accuracy. The default parameters for each classifier in Keras were utilised to train on the training dataset and evaluated. Based on the process dynamics for each key component, it was anticipated that different classification algorithms would perform differently under different levels of training data quality and degradation scenarios. A simple modal framework was employed to develop a weighted ensemble classification FDI model, where the highest three occurring classification algorithms were used. The second data-driven approach was an FDI classifier based on a neural network architecture that used a recurrent neural network (RNN) model comprising a single RNN layer with 150 neurons followed by a dense layer and a sigmoid activation function (see Figure 7 below). The model was compiled with binary cross-entropy loss and the Nadam optimiser. Early stopping was then employed to prevent overfitting during the training of the model. The performance of the RNN model was also evaluated using the test degradation data recorded in Table 4.

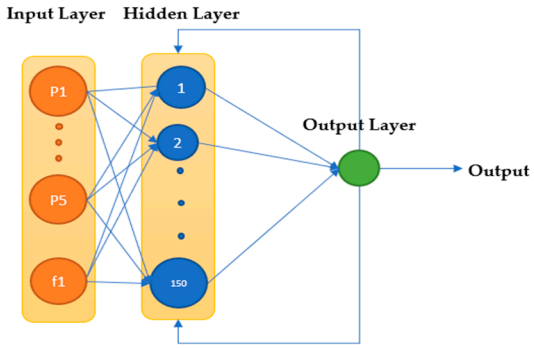

Figure 7. Recurrent neural network architecture used for each component in the water distribution system.

2.4. FDI Benchmarking Process

Figure 8 below shows a three-stage procedure for benchmarking the FDI algorithms described in Section 2.3 above. The process begins with data capture and manipulation, where the measurement system analysis process described in Section 2.1 is used to determine the quality of data of different training datasets. The function $f_1(x)$ is a logic function used to select a dataset D_1 with a specific level of quality to train the FDI algorithms in the model development and testing stage. In the statistical process control, the ensemble and neural network models are trained simultaneously, tagged as models $M_1, M_2,$ and M_3, respectively, and stored in a fault detection and isolation model repository. The function $f_2(x)$ is then used to determine the proportion of accurate predictions of test degradation scenario data by $M_1, M_2,$ and M_3.

3. Results and Potential IIoT instantiation

3.1. Results and Findings

3.1.1. Measurement System Analysis

Figure 9 below shows the relative pooled standard deviation (RPSD) for the historical process data recorded over the four-week period. In the case of the degradation scenarios, the maximum standard deviation of data recorded relative to the mean was 950 rpm for

the clogged filter failure condition scenario, with the minimum occurring at 700 rpm for a leaking pipe scenario. The healthy condition scenario showed near zero RPSD for the varying pump speeds over the four-week period. Even though Figure 9 provides insights into the variation in the data, it was difficult to determine the source of the variation, whether it was from the measurement system or the process itself. Figure 10 shows the nature of data recorded by the measurement system of the water distribution testbed. Over the four weeks, the data quality scores as defined in this paper showed data characterising the conditions of the pump with the lowest data-quality score at 0.986 and that of the valve at 0.9942. The quality of one of the synthetic datasets generated is also shown in Figure 10, with data characterising the conditions of the filter with the lowest data-quality score at 0.76 and that of the valve at 0.8.

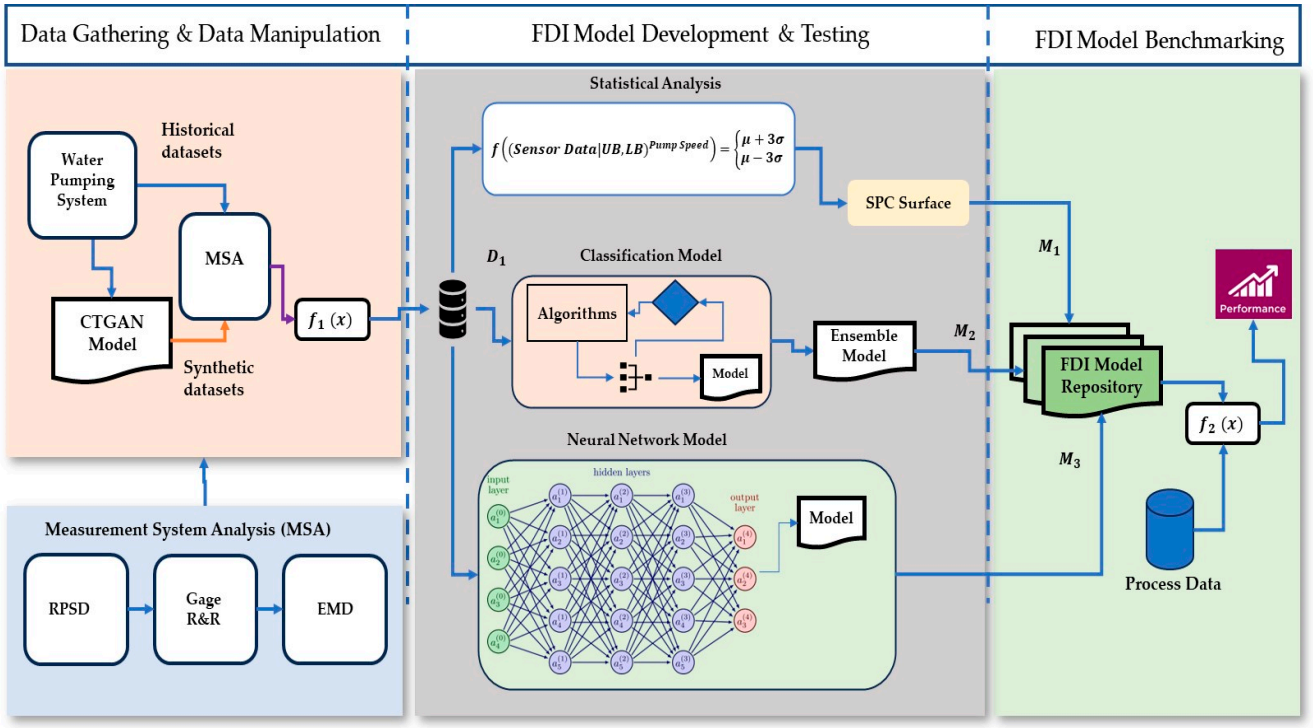

Figure 8. Process for benchmarking the FDI algorithms.

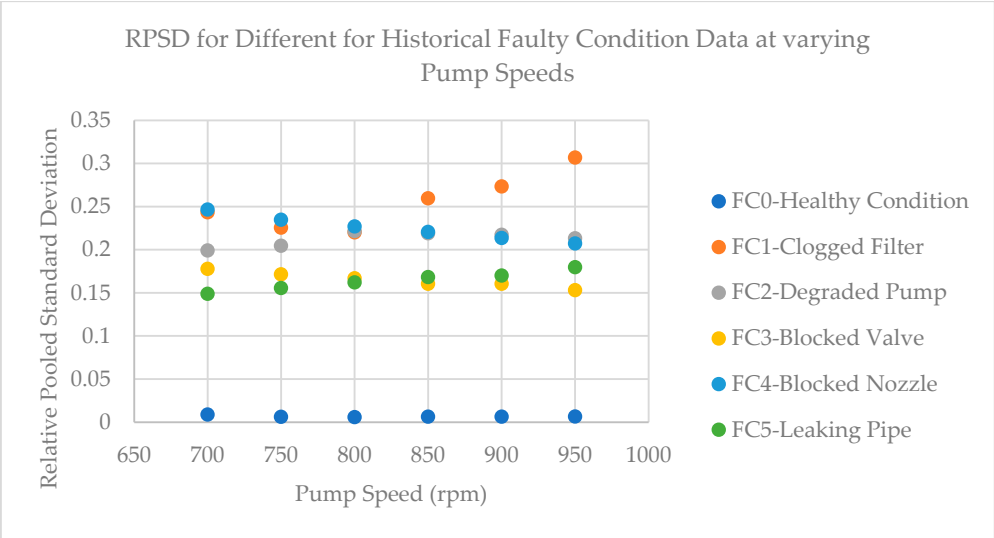

Figure 9. RPSD for different faulty condition data at varying pump speeds.

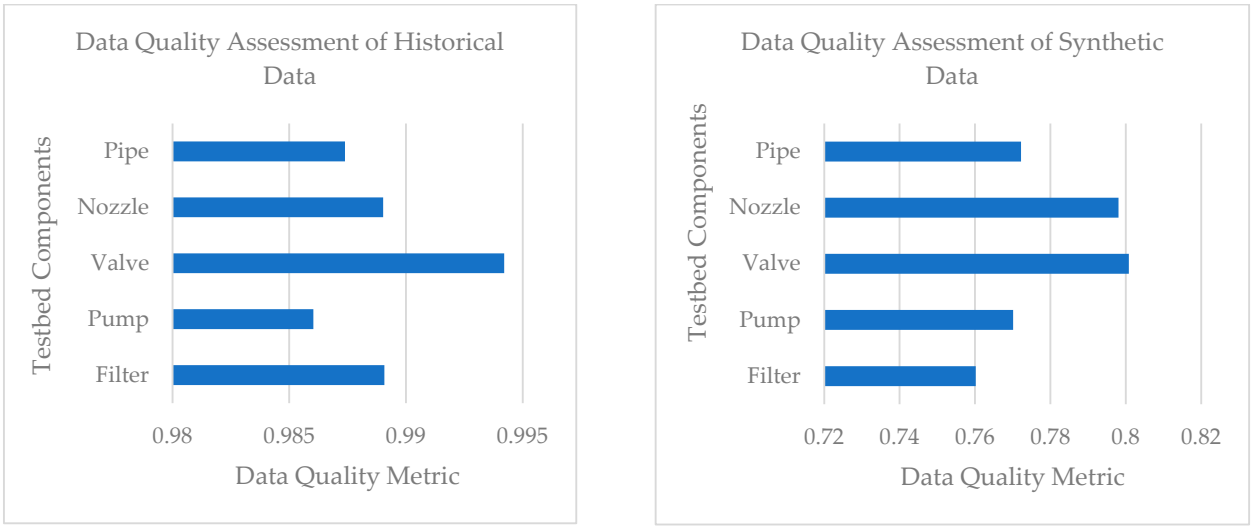

Figure 10. Data-quality assessment.

3.1.2. The Nexus between Data Quality and FDI Model Performance

The relationship between the quality of data defined by the MSA process and its impact on the development of fault detection and isolation algorithms varies depending on the nature of the principles that underpin the mathematical foundations and development process of those algorithms. The performance of the models trained with both historical and synthetic datasets is presented below, with Appendix A.2 (Tables A1 and A2) showing the best classification algorithm for each component. Figure 11 shows a radar plot of the data-quality score for every component and its impact on test degradation data T13. In the case of the SPC model, the impact of the synthetic data with lower data quality compared to the historical dataset resulted in an increase in model performance, where the synthetic dataset produced a mean accuracy for all components of 85.55% and the historical dataset producing a mean accuracy of 49.7%. The same pattern was seen in T17 as well, where the performance of the data-driven models decreased with a decrease in data quality (see Figure 12). Tables 5 and 6 show a summary of the performance of the FDI algorithms using historical and synthetic data respectively. Figures 13 and 14 also show the average performance of the FDI models in the rest of the test degradation scenarios. It was observed that there was no direct relationship between data quality as defined and the performance of the SPC FDI algorithms in the detection MCD scenarios.

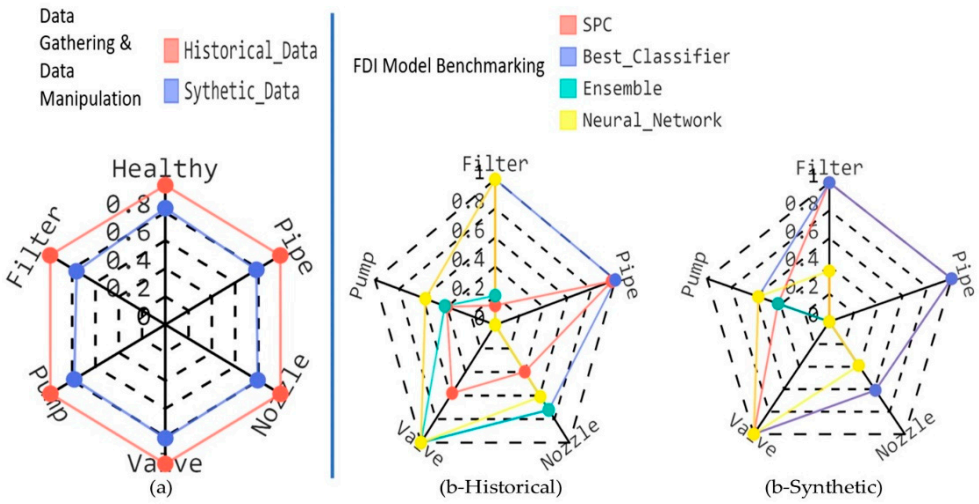

Figure 11. Impact of data quality (a) on performance of FDI models (b-Historical) and (b-Synthetic) for T13.

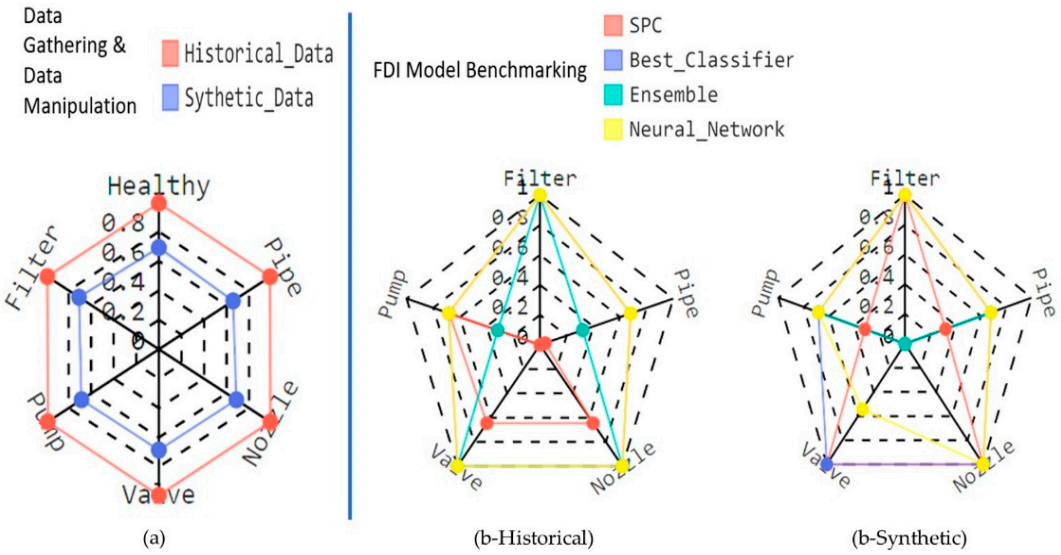

Figure 12. Impact of data quality (a) on performance of FDI models (b-Historical) and (b-Synthetic) for T17.

Table 5. Summary of average performance of FDI models for T13.

Historical Dataset with Average Data Quality of 0.99				
Model	SPC	Best Classifier	Ensemble	Neural Network
Mean accuracy (%)	49.7	86.10	46.84	63.81
Synthetic Dataset with Average Data Quality of 0.794				
Model	SPC	Best Classifier	Ensemble	Neural Network
Mean accuracy (%)	80.55	83.82	16.18	46.71

Figure 13. Performance of FDI models using historical dataset for test degradation scenarios.

Table 6. Summary of average performance of FDI models for T17.

Historical Dataset with Average Data Quality of 0.99				
Model	SPC	Best Classifier	Ensemble	Neural Network
Mean accuracy (%)	40.18	87.28	72.72	87.27
Synthetic Dataset with Average Data Quality of 0.69				
Model	SPC	Best Classifier	Ensemble	Neural Network
Mean accuracy (%)	72.72	87.27	27.27	78.12

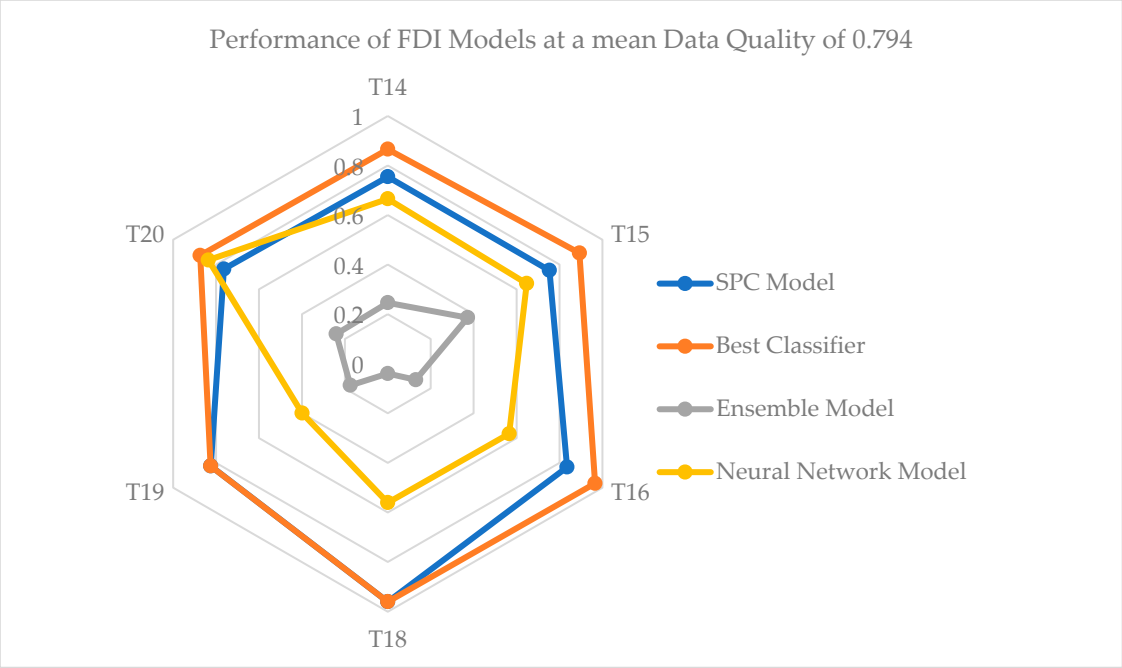

Figure 14. Performance of FDI models using synthetic dataset for test degradation scenarios.

3.1.3. Discussion

The results above present interesting model performance results as a consequence of data quality. Figure 15 below shows why there was no direct relationship between the SPC FDI performance algorithm and the quality of the datasets defined in this paper. Assuming T_1, T_2 are the lower and upper bounds, respectively, for the description of a process condition on a system; x_1, x_2 are the process parameters at time t_1, t_2, respectively, and S_1, S_2 are the lower and upper bounds, respectively, for datasets with low data quality, the SPC FDI algorithm returned a true prediction only when the conditions in cases 1, 2 and 3 were met regardless of the level of quality in the data recorded by the measurement system (S: synthetic data, A: actual data, P: performance). Case 4 is the only scenario where a true prediction was not observed by the SPC FDI algorithm. That explains why in Figures 13 and 14 the SPC model developed with a dataset with an average quality of 0.794 performed better than the data from the testbed's measurement system, with a data quality of 0.99. From the analysis of the performance of the SPC FDI model on the rest of the test dataset (see Figures 13 and 14) for the various multi-component degradation scenarios, synthetic data could be used to augment the performance of the SPC model for better performance, and it is worth investigating the requirements for an augmented SPC model with synthetic data.

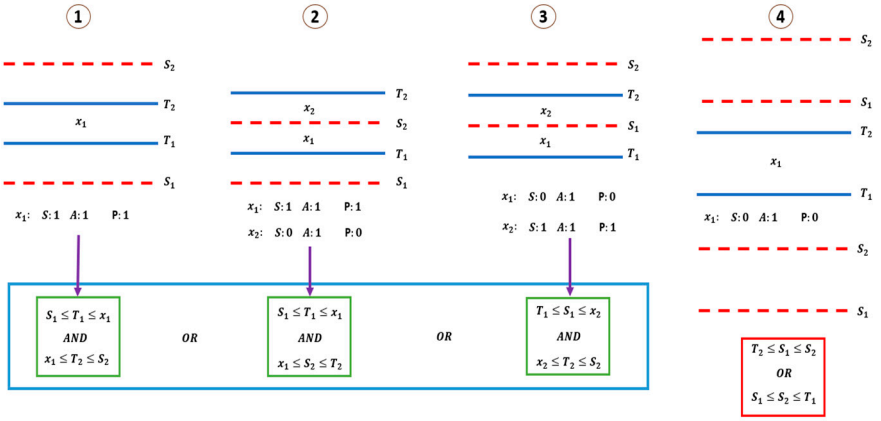

Figure 15. Relation between data distribution and SPC FDI prediction (True: 1, False: 0).

On the other hand, the data-driven methods in most cases required datasets with high data-quality levels for better model performance, as seen in Figures 11 and 12 as well as Figures 13 and 14. The neural network model showed adaptive behaviour for model prediction due to a neural network's ability to approximate system dynamics with effective architecture. Its performance did not generally decrease with a decline in data quality, unlike what was seen in the ensemble model. Implementing the best classifier approach would be impractical from an operational standpoint, as the best classifiers for a specific case of multi-component degradation as the best classification algorithms might not be known and would require simulating and testing all possible MCD combinations. For a large system, this becomes untenable because of resource constraints.

3.2. Potential IIoT Instantiation of the MSA Process

For a system $\dot{X}_iOC_i \ldots\ldots\ldots\ldots \dot{X}_NOC_N$, where $\dot{X}_i$ is a state the system can assume for the operating condition OC_i, a database $DB_{\dot{X}_iOC_i}$ with tables $T_{\dot{X}_iOC_{i1}} \ldots\ldots\ldots T_{\dot{X}_NOC_{NN}}$ can be created with replications R_i. The database with the tables is then stored on the edge and the MSA function is applied to them. Using the benchmarking process defined in Figure 8, a data-quality parameter q can be defined for a specific FDI algorithm, where the output of the MSA process is used to determine the data to be cached for training and testing. Figure 16 below shows how the MSA process can be implemented on digital network architecture. This allows for tailoring the training and test datasets to specific fault detection and isolation algorithms based on the impact of data quality on model performance.

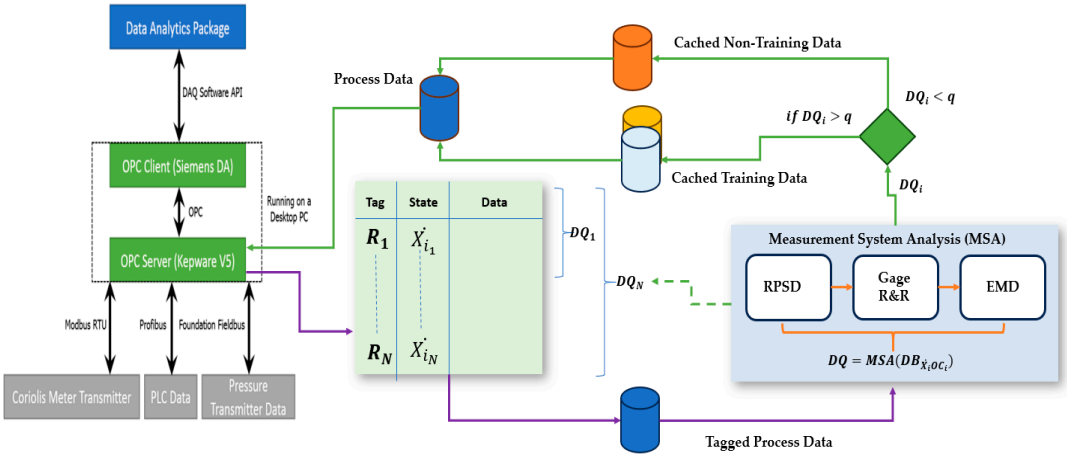

Figure 16. IIoT implementation of the measurement system analysis (MSA) process.

4. Concluding Remarks

In this research, we explored the relationship between data quality, as defined by a measurement system analysis (MSA) process, and the performance of fault detection and isolation (FDI) algorithms. Our findings show the nexus between data quality and the reliability of FDI models in the context of complex systems undergoing multi-component degradation scenarios. Our analysis of historical process data, using metrics such as the relative pooled standard deviation (RPSD) and data-quality scores for individual components, provides valuable insights. It was observed that data quality is not a singular, universal determinant of FDI model performance. Although higher data quality appears to benefit data-driven FDI models, it was interesting that, in certain cases, synthetic data with lower data quality could enhance the performance of statistical process control (SPC) models. A better understanding of the impact of data quality on FDI algorithms for PHM applications would provide insights into the deployment of specific prognostic algorithms for a system such as a city-wide water distribution system. Also, from an operational and maintenance perspective, better FDI capabilities increase the overall water distribution system's uptime, as well as the likelihood of developing, for the critical components of such a system, accurate potential failure–functional failure (P-F) degradation curves. The knowledge capture related to P-F curves can also help with improvements in the operations and optimisation of maintenance activities.

Furthermore, this non-linear relationship between data quality and FDI performance, as shown in our results and further in our discussion, highlights the need for a targeted approach in the design and implementation of FDI algorithms for multi-component degradation scenarios. The potential instantiation of the MSA process to optimise data quality for FDI models presents an avenue for future research and application. By creating databases with tailored data-quality parameters, training and testing datasets can be optimised to suit the requirements of specific FDI algorithms for multi-component degradation scenarios. This data optimisation approach using the MSA process provides a practical means to enhance the robustness and reliability of FDI systems in real-world industrial applications.

In conclusion, this research highlights the multidimensional nature of data quality's impact on FDI model performance. Leveraging MSA principles and suitable network architectures, we can open doors to a new era of precision and effectiveness in fault detection and isolation within complex systems such as a city-wide water distribution system undergoing multi-component degradation. This work contributes to the ongoing evolution of data-driven technologies to enable the monitoring of smart infrastructure. The ability of a city-wide water distribution system to meet the needs of a growing city population requires that assets in the system be monitored with robust FDI algorithms with a clear understanding of the impact of data quality in the FDI model development process.

Author Contributions: Conceptualisation, A.K.B. and O.N.; methodology, A.K.B.; software, A.K.B.; validation, A.K.B., O.N. and D.M.; formal analysis, A.K.B. and O.N.; investigation, A.K.B.; resources, A.K.B. and O.N.; data curation, A.K.B.; writing—original draft preparation, A.K.B.; writing—review and editing, A.K.B., O.N., A.C. and D.M.; visualisation, A.K.B.; supervision, O.N., A.C. and D.M.; project administration, A.K.B. and O.N.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for this research contained or presented in the tables in the article.

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

Appendix A

Appendix A.1. Test Degradation Scenarios

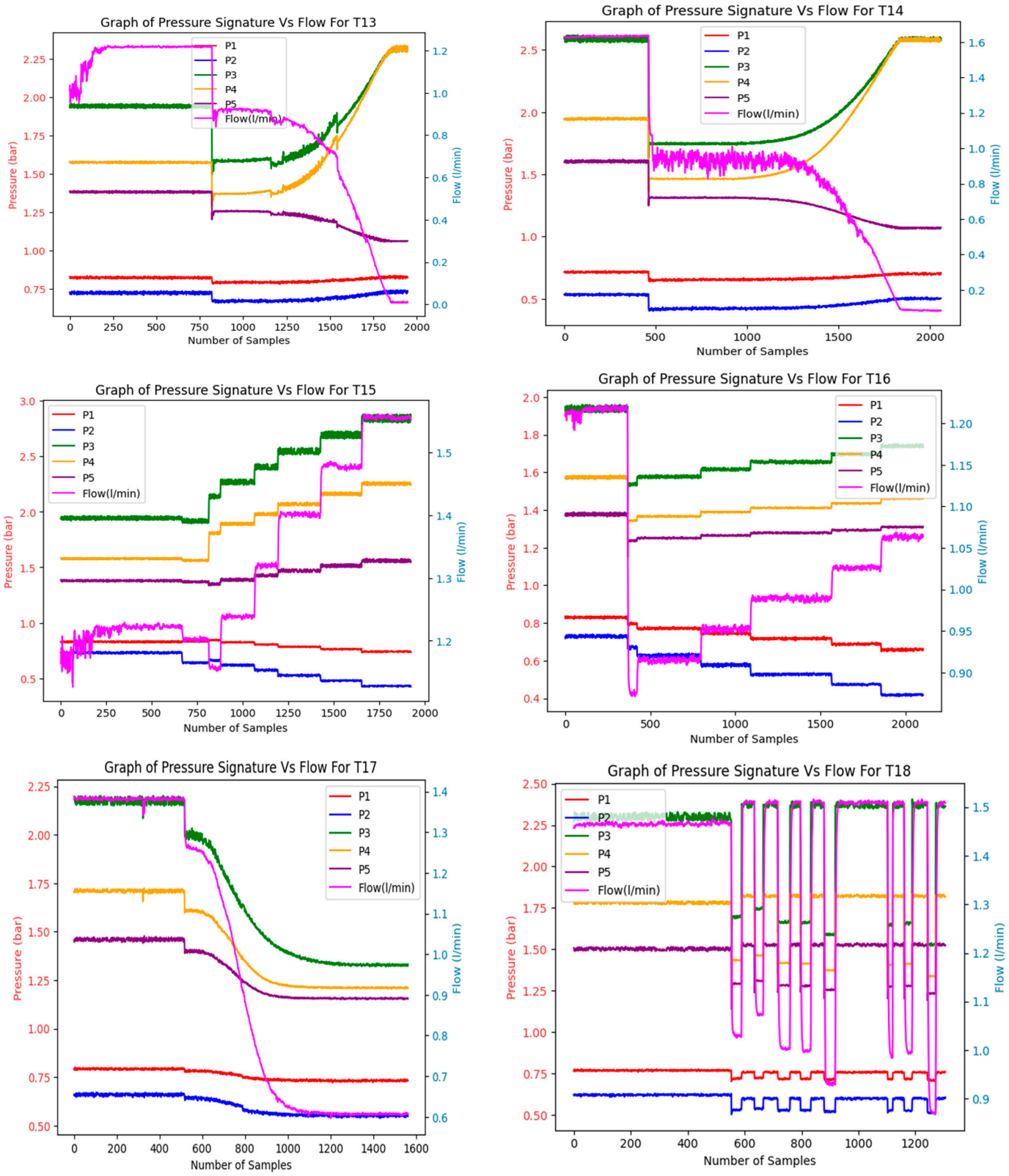

Figure A1. *Cont.*

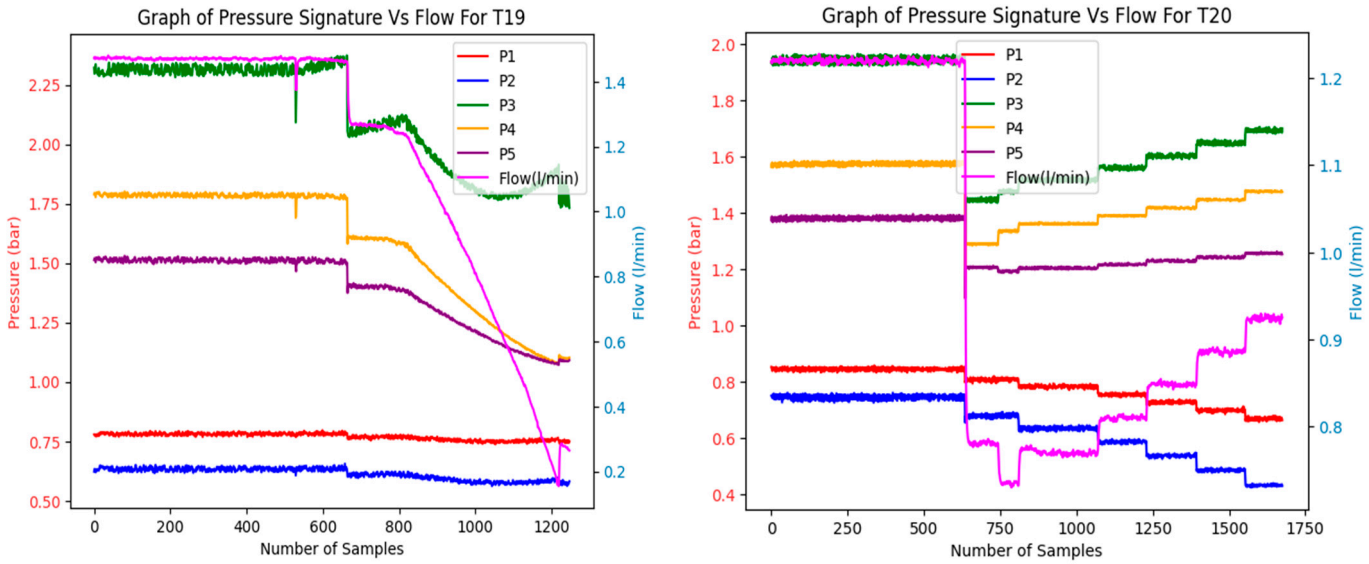

Figure A1. Test Degradation Scenarios.

Appendix A.2. Best Classification Algorithm for Each Component Ungergoing Component Degradation

Table A1. Historical dataset.

Test	Filter	Pump	Valve	Nozzle	Pipe
T13	Logistic regression	Logistic regression	Logistic regression	Logistic regression	Decision tree classifier
T14	Decision tree classifier	SVC	Logistic regression	Decision tree classifier	Logistic regression
T15	Decision tree classifier	SVC	Logistic regression	Decision tree classifier	Gaussian NB
T16	SVC	Decision tree classifier	Gaussian NB	SVC	SVC
T17	Logistic regression	Decision tree classifier	Gaussian NB	Logistic regression	K neighbours classifier
T18	SVC	Decision tree classifier	Decision tree classifier	Logistic regression	Decision tree classifier
T19	Logistic regression	Decision tree classifier	Logistic regression	Decision tree classifier	SVC
T20	SVC	Logistic regression	Logistic regression	Decision tree classifier	SVC

Table A2. Synthetic dataset.

Test	Filter	Pump	Valve	Nozzle	Pipe
T13	Logistic regression	Logistic regression	Decision tree classifier	Logistic regression	Logistic regression
T14	Decision tree classifier	Decision tree classifier	Gaussian NB	Decision tree classifier	Gaussian NB
T15	Decision tree classifier	Decision tree classifier	Decision tree classifier	Logistic regression	Decision tree classifier
T16	SVC	Logistic regression	Decision tree classifier	Gaussian NB	Gaussian NB
T17	Logistic regression	Logistic regression	Decision tree regressor	Decision tree regressor	Logistic regression
T18	SVC	Logistic regression	Decision tree classifier	Decision tree classifier	Decision tree classifier
T19	Logistic regression	Decision tree classifier	Decision tree classifier	Decision tree classifier	Decision tree classifier
T20	SVC	Logistic regression	Logistic regression	Logistic regression	Logistic regression

References

1. Fuller, A.; Fan, Z.; Day, C. Digital Twin: Enabling Technologies, Challenges and Open Research. *IEEE Access* **2020**, *8*, 108952–108971. [CrossRef]
2. Tao, F.; Qi, Q.; Wang, L.; Nee, A.Y.C. Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: Correlation and comparison. *Engineering* **2019**, *5*, 653–661. [CrossRef]
3. Lu, Y.; Liu, C.; Kevin, I.; Huang, H.; Xu, X. Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robot. Comput. Integr. Manuf.* **2020**, *61*, 101837. [CrossRef]
4. Malakuti, S.; Grüner, S. Architectural aspects of digital twins in IIOT systems. In Proceedings of the 12th European Conference on Software Architecture: Companion Proceedings, Madrid, Spain, 24–28 September 2018; pp. 1–2. [CrossRef]
5. Kuts, V.; Modoni, G.E.; Otto, T.; Sacco, M.; Tahemma, T.; Bondarenko, Y.; Wang, R. Synchronizing physical factory and its digital twin through an IIOT middleware: A case study. *Proc. Est. Acad. Sci.* **2019**, *68*, 364. [CrossRef]

6. Souza, V.; Cruz, R.; Silva, W.; Lins, S.; Lucena, V. A digital twin architecture based on the industrial internet of things technologies. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; pp. 1–2.
7. Delgado, J.M.D.; Odeyele, L. Digital Twins for the built environment: Learning from conceptual and process models in manufacturing. *Adv. Eng. Inform.* **2021**, *49*, 101332. [CrossRef]
8. Yao, J.F.; Yang, Y.; Wang, X.C.; Zhang, X.P. Systematic review of digital twin technology and applications. *Vis. Comput. Ind. Biomed. Art* **2023**, *6*, 10. [CrossRef] [PubMed]
9. Aghashahi, M.; Sela, L.; Banks, M.K. Benchmarking dataset for leak detection and localization in water distribution systems. *Data Brief* **2023**, *48*, 109148. [CrossRef] [PubMed]
10. Fernandes, M.; Corchado, J.M.; Marreiros, G. Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: A systematic literature review. *Appl. Intell.* **2022**, *52*, 14246–14280. [CrossRef]
11. Han, H.; Gu, B.; Hong, Y.; Kang, J. Automated FDD of multiple-simultaneous faults (MSF) and the application to building chillers. *Energy Build.* **2011**, *43*, 2524–2532. [CrossRef]
12. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
13. Yang, J.; Guo, Y.; Zhao, W. Long short-term memory neural network based fault detection and isolation for electro-mechanical actuators. *Neurocomputing* **2019**, *360*, 85–96. [CrossRef]
14. Khalid, S.; Khalil, T.; Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. In Proceedings of the 2014 Science and Information Conference, London, UK, 27–29 August 2014; IEEE: New York, NY, USA; pp. 372–378.
15. Zhang, L.; Lin, J.; Liu, B.; Zhang, Z.; Yan, X.; Wei, M. A review on deep learning applications in prognostics and health management. *IEEE Access* **2019**, *7*, 162415–162438. [CrossRef]
16. Burdick, R.K.; Borror, C.M.; Montgomery, D.C. A Review of Methods for Measurement Systems Capability Analysis. *J. Qual. Technol.* **2018**, *35*, 342–354. [CrossRef]
17. Dina, A.S.; Siddique, A.B.; Manivannan, D. Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks. *IEEE Access* **2022**, *10*, 96731–96747. [CrossRef]
18. Hozo, S.P.; Djulbegovic, B.; Hozo, I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med. Res. Methodol.* **2005**, *5*, 13. [CrossRef] [PubMed]
19. Kazerouni, A.M. Design and analysis of gauge R&R studies: Making decisions based on ANOVA method. *Int. J. Ind. Manuf. Eng.* **2009**, *3*, 335–339.
20. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
21. Delouche, N.; Dersoir, B.; Schofield, A.B.; Tabuteau, H. Flow decline during pore clogging by colloidal particles. *Phys. Rev. Fluid.* **2022**, *7*, 034304. [CrossRef]
22. Wahab, A. Analytical prediction technique for internal leakage in an external gear pump. *Turbo Expo Power Land Sea Air* **2009**, *48869*, 85–92.
23. Ortega, R.; Metcalf, J.G.; Patton, C. Improving Operations of Inoperable Valves. In *Infrastructure's Hidden Assets, Proceedings of the Pipelines 2009, San Diego, CA, USA, 15–19 August 2009*; American Society of Civil Engineers: Reston, VA, USA; pp. 1075–1082.
24. Zhu, B.; Liu, Q.; Zhao, D.; Ren, S.; Xu, M.; Yang, B.; Hu, B. Effect of Nozzle Blockage on Circulation Flow Rate in Up-Snorkel during the RH Degasser Process. *Steel Res. Int.* **2016**, *87*, 136–145. [CrossRef]
25. Abtew, M.A.; Kropi, S.; Hong, Y.; Pu, L. The Application of Using Statistical Process Control (SPC) Tools in Improving the Quality of a Manufacturing Process: A Case Study. *Spektrum Ind.* **2018**, *11*, 105–1142.
26. Iqball, T.; Wani, M.A. Weighted ensemble model for image classification. *Int. J. Inf. Technol.* **2023**, *15*, 557–564. [CrossRef] [PubMed]
27. Sperandei, S. Understanding logistic regression analysis. *Biochem. Medica* **2014**, *24*, 12–18. [CrossRef] [PubMed]
28. Charbuty, B.; Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
29. Alam, M.S.; Vuong, S.T. Random Forest classification for detecting android malware. In Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, Beijing, China, 20–23 August 2013; pp. 663–669. [CrossRef]
30. Bafjaish, S.S. Comparative analysis of Naive Bayesian techniques in health-related for classification task. *J. Soft Comput. Data Min.* **2020**, *1*, 1–10.
31. Kataria, A.; Singh, M.D. A review of data classification using k-nearest neighbor algorithm. *Int. J. Emerg. Technol. Adv. Eng.* **2013**, *3*, 354–360.
32. Zhang, Y. Support vector machine classification algorithm and its application. In Proceedings of the Information Computing and Applications: Third International Conference, Chengde, China, 14–16 September 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 179–186.

33. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
34. Wang, R. AdaBoost for feature selection, classification and its relation with SVM, a review. *Phys. Procedia* **2012**, *25*, 800–807. [CrossRef]