Preprints.org

Article

Not peer-reviewed version

Best Practice Data Sharing Guidelines for Wind Turbine Fault Detection Model Evaluation

Sarah Barber * , Unai Izagirre Aizpitarte , Oscar Serradilla , Jon Olaizola Alberdi , Ekhi Zugasti Uriguen , Jose Aizpurua , Ali Eftekhari Milani , Frank Sehnke , Yoshiaki Sakagami , Charles Henderson

Posted Date: 14 March 2023

doi: 10.20944/preprints202303.0239.v1

Keywords: Wind energy; Data sharing; Best practice; Machine learning; Model evaluation

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

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.

Article

Best Practice Data Sharing Guidelines for Wind Turbine Fault Detection Model Evaluation

Sarah Barber [1,*], Unai Izagirre [2], Oscar Serradilla [2], Jon Olaizola [2], Ekhi Zugasti [2], Jose I. Aizpurua [2,3], Ali Eftekhari Milani [4], Frank Sehnke [5], Yoshiaki Sakagami [6] and Charles Henderson [7]

1 Eastern Switzerland University of Applied Sciences, Oberseestrasse 10, 8640 Rapperswil, Switzerland; sarah.barber@ost.ch
2 Mondragon University, Electronics & Computer Science Department, Goiru , Arrasate, Spain
3 Ikerbasque, Basque Foundation for Science, Bilbao, Spain
4 TU Delft, Delft, Netherlands
5 Center for Solar Energy and Hydrogen Research - ZSW, Stuttgart, Germany
6 Federal Institute of Santa Catarina, Florianópolis, Brazil
7 Stacker Group, Charlottesville, Virginia, USA
* Correspondence: sarah.barber@ost.ch

Abstract: The digital era offers many opportunities to the wind energy industry and research community. Digitalisation is one of the key drivers for reducing costs and risks over the whole wind energy project life cycle. One of the largest challenges in successfully implementing digitalisation is the lack of data sharing and collaboration between organisations in the sector. In order to overcome this challenge, a new collaboration method called WeDoWind was developed in recent work. The main innovation of this method is the way it creates tangible incentives to motivate and empower different types of people from all over the world to actually share data and knowledge in practice. In this present paper, the challenges related to comparing and evaluating different SCADA data based wind turbine fault detection models are investigated by carrying out a new case study, the "WinJi Gearbox Fault Detection Challenge", based on the WeDoWind Method. Six new solutions were submitted to the challenge, and a comparison and evaluation of the results show that, in general, some of the approaches (Particle Swarm Optimisation algorithm for constructing health indicators, performance monitoring using Deep Neural Networks, Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor and Time-to-failure prediction using Random Forest Regression) appear to have a high potential to reach the goals of the Challenge. However, there are a number of concrete things that would have to have been done by the Challenge providers and the Challenge moderators in order to ensure success. This includes enabling access to more details of the different failure types, access to multiple data sets from more wind turbines experiencing gearbox failure, provision of a model or rule relating fault detection times or a remaining useful lifetime to the estimated costs for repairs, replacements and inspections, provision of a clear strategy for training and test periods in advance, as well as provision of a pre-defined template or requirements for the results. These learning outcomes are used directly to define a set of best practice data sharing guidelines for wind turbine fault detection model evaluation. They can be used by the sector in order to improve model evaluation and data sharing in the future.

Keywords: wind energy; data sharing; best practice; machine learning; model evaluation

1. Introduction

1.1. The opportunities and challenges of digitalisation

The digital era offers many opportunities to the wind energy industry and research community. Digitalisation, defined as "the organisational and industry-wide use of data and digital technologies to

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

improve efficiency, create insights, and develop products and services" [1], is one of the key drivers for reducing costs and risks over the whole wind energy project life cycle (https://windeurope.org/intelligence-platform/product/wind-energy-digitalisation-towards-2030/; accessed on 3 August 2022). The main opportunities relate to new processes and business models brought about by the availability - and innovative usage - of large amounts of data at all phases of the wind project life-cycle, including wind farm layout design, construction, commissioning, maintenance and operations. These can include, for example, digital twins, predictive maintenance, drones and expert systems, which can all contribute to reducing maintenance costs, increasing the amount of energy delivered as well as increasing efficiency.

However, a successful exploitation of these opportunities raises a number of challenges, which were recently thoroughly examined in [1]. This resulted in the definition of three "Grand Challenges" of wind energy digitalisation: (1) Creating findable, accessible, interoperable and reusable (FAIR) data frameworks [**?**]; (2) Connecting people and data to foster innovation; (3) Enabling collaboration and competition between organisations.

Solutions to overcoming these Grand Challenges have already been investigated to some extent, as described in [2]. Efforts include the Sharewind metadata registry [3], the IEA Wind Task 43 Metadata Challenge [4], digital platforms for collecting data from interactions with stakeholders [5], data marketplaces such as the IntelStor Market Intelligence Ecosystem (https://www.intelstor.com/; accessed on 3 August 2022), the EDP Open Data Platform (https://opendata.edp.com/; accessed on 3 August 2022), open-source tools including the Brightdata app (https://www.brightwindanalysis.com/brightdata/; accessed on 3 August 2022), OpenOA (https://github.com/NREL/OpenOA; accessed on 3 August 2022) and the Data Science for Wind Energy R Library (https://github.com/TAMU-AML/DSWE-Package/; accessed on 3 August 2022), as well as collaborative innovation processes [5,6]. However, none of these initiatives overcome the three Grand Challenges in a holistic manner, nor do they specifically enable cooperation between organisations or connect people and data.

In order to close this gap, a new collaboration method called WeDoWind was developed in recent work by the present lead author [2], as described below. The main innovation of this method is the way it creates tangible incentives to motivate and empower different types of people from all over the world to actually share data and knowledge in practice.

1.2. The WeDoWind Method

The idea behind the WeDoWind Method is summarised in the value creation process in Figure 1. It brings together WeDoWind Challenge Providers, such as wind farm owner/operators, measurement service providers, hardware suppliers and asset managers, who provide data and WeDoWind Challenges, with Solution Developers, such as researchers, students and data scientists, who use the provided data to solve the WeDoWind Challenges. The data required in order to solve a particular WeDoWind Challenge are provided by the WeDoWind Challenge Providers under the confidentiality conditions they specify. This can include only allowing specific people to access their space, requiring them to sign agreements or preparing the data so that it is anonymous or normalised. A WeDoWind Challenge is defined as a fixed problem with a motivation, goal, expected outcome and deadline.

The WeDoWind Challenge Providers are motivated to participate because they receive diverse state-of-the-art solutions to their WeDoWind Challenges from international teams, increased visibility and contact with the research community. The Solution Developers benefit by getting access to relevant WeDoWind Challenges and data from the industry and getting visibility within a vibrant international research and teaching community. The digital platform allows all the data, code, meeting documentation, questions and discussions to be documented in one central location, meaning that it can be built upon in future WeDoWind Challenges and easily accessed by a wide range of people all over the world. The resulting community of active and passive challenge participants from academia and industry interested in sharing data, code and learning from each other forms a thriving and open digital wind energy ecosystem.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

Ultimately, the method allows the industry to adopt new approaches developed by academia much faster than is currently being done. This is because it allows the accumulation of data sets from multiple data providers in a unified format, providing a more realistic environment for developing and benchmarking new models. Additionally, it enables testing on how well the developed models scale to a large database, an integral requirement in practice. Finally, it can enforce transparency and robustness requirements for challenge solutions.

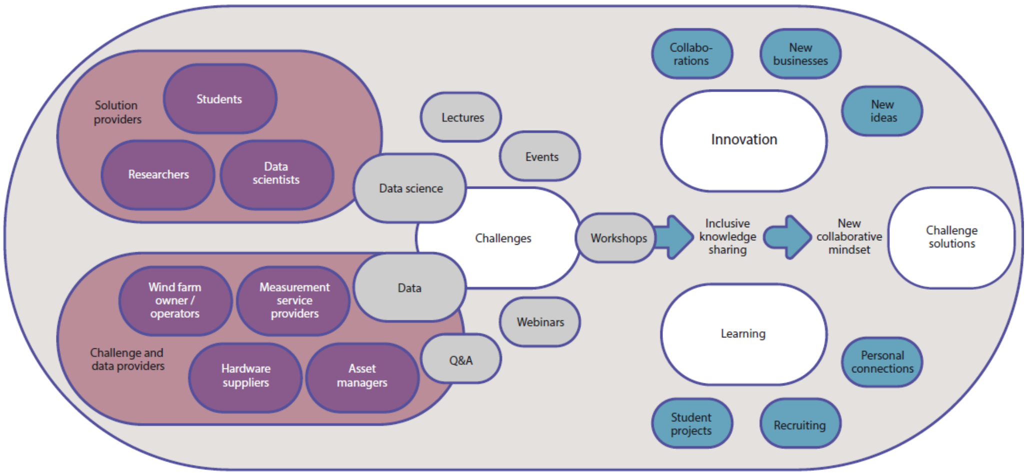

Figure 1. The value creation process of the WeDoWind Method.

In previous work, the method was tested on a case study, the "EDP Wind Turbine Fault Detection Challenge" [2]. The aim of this WeDoWind Challenge was "to identify failures in five of the major wind turbine components and advise an intervention to the wind farm operators in order to reduce corrective maintenance costs". The WeDoWind Method was applied by creating a dedicated space on the digital platform. The entire WeDoWind Challenge process was planned and coordinated by the method developers at the Eastern Switzerland University of Applied Sciences (OST), which involved the acquisition of participants, moderating and documenting workshops, offering support using the digital platform, sending regular email updates, providing a downloadable docker for beginners and evaluating and disseminating the results. Six solutions using Normal Behaviour Models, Combined Local Minimum Spanning Tree and Cumulative Sum of Multivariate Time Series Data, Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor, Normal Behaviour Model with Lagged Inputs, Canonical Correlation Analysis and Kernel Change-Point Detection were submitted to this WeDoWind Challenge. Evaluation of the results showed several advantages and disadvantages of the different solutions. Two of the solutions performed significantly better than EDP's existing method in terms of total prediction costs.

The case study demonstrated that the WeDoWind Method is a promising solution for enhancing collaboration and data sharing in wind energy. It provided a number of benefits for both WeDoWind Challenge and Solution Providers, including access to data, code, knowledge and people skills. It was found to have a high potential to be used by the wider community, both within and beyond wind energy, by providing realistic environments for developing and benchmarking new machine learning (ML) models.

It also allowed several challenges relating to the evaluation and comparison of different models to be identified. For example, the wide range of comparison metrics and Key Performance Indicators (KPIs) make it difficult to assess which metric or KPI is most suitable for a given application, especially when the heterogeneity of the models is high. As well as this, the conversion of these metrics into financial gains involves a number of different assumptions, and varying these assumptions in the EDP Challenge leads to a large variation in the results [2]. Finally, the quality of the evaluation results was

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

thought to be dependent on the quality and quantity of provided data. These issues all require further investigation if data sharing and collaboration are to be improved in the sector.

In order to do this, multiple different WeDoWind Challenges need to be introduced and run, enabling both the accumulation of data sets from multiple data providers as well as the creation of an active community of people learning and sharing with each other. This effort is furthered in the current paper, with a focus on wind turbine fault detection based on SCADA data.

1.3. Wind turbine fault detection based on SCADA data

The topic of wind turbine fault detection based on SCADA data is chosen for this work because it is an accessible topic for the industry and an interesting research topic for academia. SCADA data is more readily available than higher frequency data sets, and additional measurement equipment such as high-frequency Condition Monitoring Systems are not required.

Although the usage of SCADA data for wind turbine monitoring is attractive, its application is limited by the low resolution of the data (usually ten minute averages). Therefore fault detection methods based on SCADA data usually target secondary effects of the fault. Temperature monitoring is well-suited for detecting malfunctions in the components along the drive train, which account for the majority of turbine downtime [7]. [8] was among the first to apply the approach in the wind domain and prove its feasibility. Many publications with successful early detection of malfunctions followed, e.g., [9–14].

A recent review of the application of machine learning (ML) models for condition monitoring in wind energy shows that most models use SCADA or simulated data, and almost two-thirds of the methods use classification, the rest relying on regression. Neural networks, support vector machines and decision trees are most commonly used [15]. A wide range of ML methods have proven to be able to detect developing malfunctions at an early stage, often months before they resulted in costly component failures (see, e.g., [8,12,13,16].

Furthermore, feature selection - a process of selecting variables that are significantly related to the outcome that the model is predicting - can be used to improve the success of condition monitoring based on SCADA data. For example, [17] used deep learning to select the relevant variables related to transmission bearing temperature before analysing SCADA data. As well as this, [18] propose an adaptive neuro fuzzy inference system to estimate the remaining useful life (RUL) of bearings based on vibration data and feature extraction. Other methods for predicting the RUL of wind turbine components include a feature fusion model based health indicator construction and self-constraint state-space estimator for bearings [19] and a novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model [20].

In the recent EDP Challenge mentioned in Section 1.2 above, the most promising models were found to be a Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor (WHC-LOF) and a Combined Local Minimum Spanning Tree and Cumulative Sum of Multivariate Time Series Data (LoMST-CUSUM). The WHC-LOF solution combines the Ward Hierarchical Clustering [21], which identifies and removes anomalous clusters, with the Novelty Detection with Local Outlier Factor (LOF) [22], which is used to detect the outliers associated with the failures of the wind turbine. The LoMST-CUSUM solutions works in three stages as described in [23]. First, it establishes a so-called Minimum Spanning Tree (MST) using all data points. Second, it isolates the cluster anomalies by removing the links of the global MST one by one. Third, it repeats the second step to identify point-wise anomalies. At the end, an outlier score is assigned to each of the data points, indicating the anomaly level of the point. Cumulative sum (CUSUM) is a memory-type control chart that works by accumulating consecutive sample points over time to monitor changes in the process. CUSUM-based approaches have been used in wind turbine monitoring in combination with ML plots [24,25].

Despite this progress, no particular model has yet been established as being optimal for SCADA data based fault detection. Furthermore, no comparison of the type, length and quality of data used in

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

previous studies has been carried out systematically. Finally, the optimal comparison metric or KPI is not clear due to the high heterogeneity of models. However, various recent studies have concluded that more data should be made available and comparisons with a wide range of different data should be carried out in order to do this, e.g. [2,26,27]. We aim to contribute to this gap in the present work.

1.4. Goals of the present study

The goals of this work are to (1) investigate the challenges related to comparing and evaluating different SCADA data based wind turbine fault detection models, and (2) to develop a set of best practice data sharing guidelines to improve model evaluation in the future. As the digital era progresses, data sharing and model evaluation are both going to grow, and therefore a thorough understanding of how to do these tasks effectively is becoming more and more important.

In order to achieve these goals, a new case study, the "WinJi Gearbox Fault Detection Challenge", was carried out based on the WeDoWind Method. The case study is introduced in Section 2 and the submitted solutions are described and evaluated in Section 3. This includes a discussion of the difficulties related to comparing and evaluating the models at the end of the section. After that, the best practice guidelines are introduced in Section 4. Finally, the conclusions are drawn in Section 5.

2. Case study description

2.1. The WinJi Gearbox Fault Detection Challenge

The WinJi Gearbox Fault Detection Challenge was provided by the Swiss asset management software provider WinJi AG, as described below:

We are an asset management software provider for renewable portfolios, also active in predictive maintenance. We have developed several methods for early prediction of faults based on SCADA data from wind turbines, achieving horizons of several days. Our aim is to push the performance of such predictions to at least a monthly horizon. Participants should make use of the provided SCADA data in order to train, test and validate methods that will provide clear indicators of an upcoming gearbox related fault, as well as/or a horizon-based probability of the event occurring. Over two years of 10-minute SCADA are provided, for five wind turbines in the same site. Two of these turbines experience gearbox failures leading to extended downtimes and one experiences bearing issues with no downtimes. The other two are free of significant failures. The dataset includes environmental, production and condition parameters as well as error information.

The data was anonymised by WinJi in the following ways:

- The power of each wind turbine was normalised by the rated power.
- The location and type of wind turbines was not provided.
- The time stamps were all shifted by an unspecified amount of time.
- The detailed fault information was not provided. Instead, fault occurrence indicators were provided as binary indicators, *i.e.* fault or no fault.

Whilst WinJi were involved strongly in the webinars in order to help and support the participants, they did not take part in the challenge themselves. This meant that a direct comparison between WinJi's own analysis method and the new methods developed here could not be carried out. WinJi was looking to learn from the challenge in order to further their knowledge indirectly. Their current method is based on a Normal Behaviour Model and is described further in [28].

Figure 2 shows the empirical power curve and wind rose (without filtering) for one of the wind turbines at the site (WT115). It can be observed that the power curve includes different operation conditions and the main wind direction is SSW.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

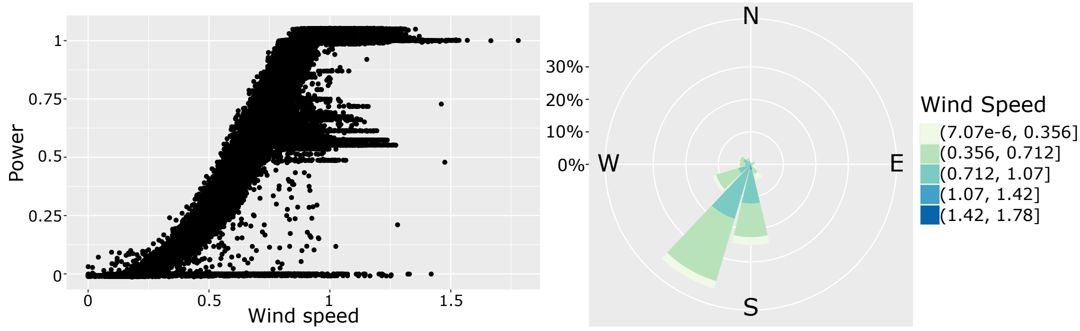

Figure 2. Power Curve and wind rose for the WT115.

2.2. The WeDoWind Method

The WeDoWind Method was applied to the case study as follows:

- A dedicated space called "WinJi Challenges" was created on the digital platform together with WinJi. The WeDoWind Challenge description, including direct links to download the data, was developed together with WinJi and posted inside this space.
- A public "call for participants" website was created with a direct link to the registration form (https://www.wedowind.ch/winji-space; accessed on 18 January 2023). This was shared within the wind energy community using social media.
- A process for allowing WinJi to decide who may participate or not was set up. This process was not meant to reduce accessibility to the WeDoWind Challenge, but instead to ensure that applicants were real people interested in the WeDoWind Challenge and not robots, bots or imposters.
- A "Getting Started Guide" to using the digital platform was created and explainer videos were recorded in order to help users interact on the platform.
- A series of online workshops were organised for the participants - a launch workshop, interim workshops every month and then a final workshop. These involved brainstorming sessions in small groups as well as question and answer sessions with WinJi. The sessions were documented on a digital whiteboard and recordings were posted in the digital space.
- Regular email updates were sent with specific questions and actions to encourage interaction. This included requests to summarise and comment on different possible methods, as well as discussions of evaluation methods.
- The space was regularly checked, cleaned and coordinated by the method developers to ensure that the information was up-to-date and understandable.
- Regular updates were communicated on social media during the challenge.

3. Submitted solutions

3.1. Description of submitted solutions

Six solutions were submitted for the challenge. The solutions are summarised in Table 1 and described in the next sub-sections below.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

Table 1. Summary of submitted solutions.

Solution	Method	Contributor	Type	Goal	Pre-Proc.	Time Res.	Training period
CAE-PSO	CAE	TUD	Unsup. Learn.	Pred. RUL	N(0, 1) Norm.	Minutes	100%
PM-DNN	DNN	ZSW	Sup. Learn.	Pred. Power Anomaly	N(0,1) Norm.	Minutes	80%-20% random Blocks
DEC-DNN	DNN	ZSW	Sup. Learn.	Pred. Error Code	N(0,1) Norm.	Minutes	80%-20% random Blocks
T-SNE	T-SNE	ZSW	Unsup. Learn.	Cluster Correlation	N(0,1) Norm.	Minutes	80%-20% random Blocks
WHC-LOF	LOF	SY	Unsup. Learn.	Pred. Anom. + RUL	WHC	Weeks	100%
TTF-RFR	RFR	MU	Sup. Learn.	TTF inference	Remove correlated features, stop instants	Hours	Leave one out (80%-20%)

3.1.1. A Convolutional Autoencoder trained with Particle Swarm Optimisation algorithm for constructing health indicators (CAE-PSO)

One of the most prevalent approaches for RUL estimation is constructing health indicators (HI), which are correlated with and act as proxies for the degradation process in the component studied. In this solution, a Convolutional Autoencoder (CAE) is proposed for unsupervised construction of health indicators from wind turbine SCADA data. The weights of the CAE are trained by a Particle Swarm Optimization (PSO) algorithm with the task of maximising the objective function $f = |\tau_{\mathrm{MK}}| - MSE$, where $|\tau_{\mathrm{MK}}| \in [0, 1]$ is the absolute value of the Mann-Kendall monotonicity score [29] of the HI constructed in the middle layer of the proposed CAE. MSE is the Mean Squared Error between the input signals and the reconstructed signals in the output layer (Figure 8). The reason for using PSO rather than backpropagation for training the CAE is the fact that monotonicity is a metric related to the entire training set as a whole and not single samples or batches of samples from the training set which the network is trained on at each iteration of the backpropagation algorithm. By maximising the monotonicity of the constructed HIs, we aim at training the CAE to extract and isolate – from the SCADA signals – the factor that is related to the degradation of the component from other factors related to the operational and environmental conditions, process noise, etc. This is according to the fact that degradation is an irreversible and monotonic process. The proposed CAE is made up of an Encoder with two Convolutional layers. The first one performs convolution only in the time domain and the second one only in the feature domain. The Decoder is symmetric to the Encoder, as shown in Figure 4.

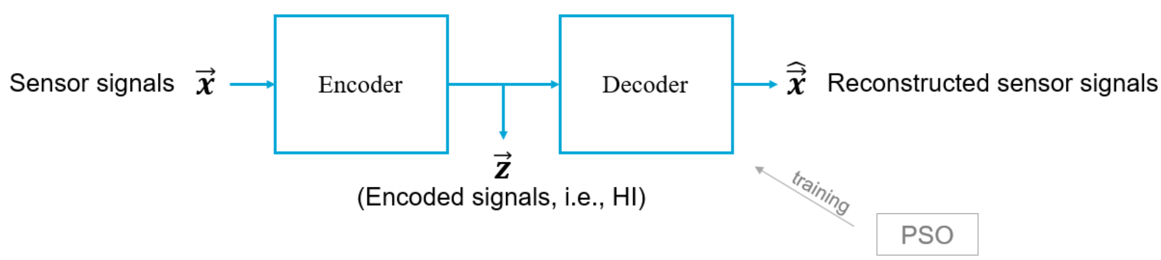

Figure 3. Architecture of the CAE

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

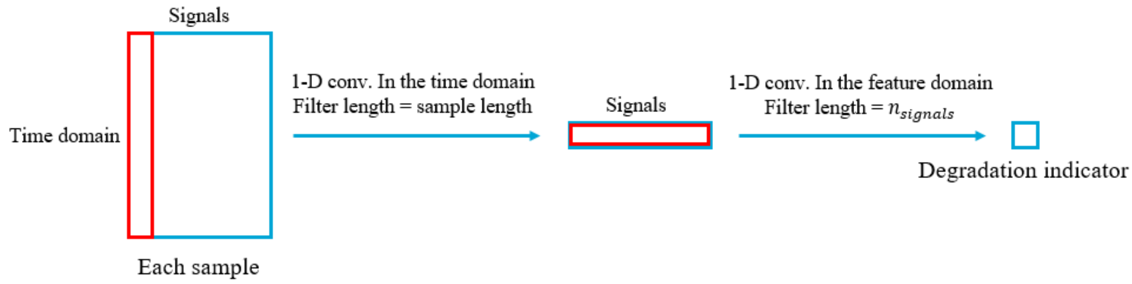

Figure 4. Architecture of the Encoder of the CAE. The Decoder is symmetric.

3.1.2. Performance monitoring using Deep Neural Networks (PM-DNN)

In this solution, the standard performance monitoring approach was investigated, using pure dense feed forward neural networks in the form of deep Multi-Layer Perceptron (MLP) [30] (otherwise known as Deep Neural Network (DNN), see Figure 5) trained with ADAM [31]. The used code is written in PyTorch [32]. The input data was normalised to a standard normal distribution (N(0,1) Norm.). Since a validation set needs to be (I) from the same distribution as the training set, but (II) also be independent of the training set, the data was split into 200 blocks and 20% of blocks were chosen randomly as validation data. This procedure is needed, since the data is a time series and therefor shows high correlation from data point to data point. Pure random selection for the validation set would violate the rule for an independent validation set.

The DNN was trained with the wind turbine power output as target. The hyper-parameters of the DNN architecture and the ADAM optimiser were optimised for the given data with Reinforcement Learning (RL). The validation set Mean Squared Error (MSE) was used as reward signal for the RL-Method Policy Gradients with Paramater-based Exploration (PGPE) [33] – we used the super symmetric, ranked version of the algorithm [34,35]. The PGPE implementation is ZSW internal python/numpy code.

For the input data we ignored the error code itself, for obvious reasons. We assumed a Gaussian error distribution of the DNN and trained another DNN on the absolute differences of the prediction of the first DNN as confidence model for the performance monitoring approach.

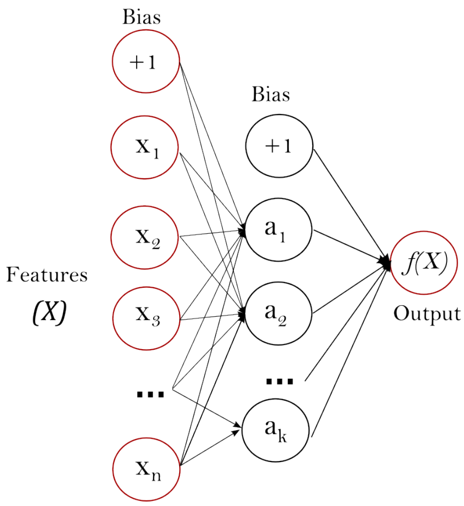

Figure 5. Standard Multi-Layer Perceptron. This kind of Neural Network was used to predict the power output as well as the confidence of the prediction. Source: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

3.1.3. Direct Error Code Prediction using DNNs (DEC-DNN)

The same kind of DNN as in the PM-DNN solution was also used to investigated the naive approach, by training directly on the error codes as binary targets. The model therefore was a standard DNN with logistic output activation function trained with ADAM and meta-optimised with PGPE. As an approximation, the output was interpreted as the probability for the occurrence of an error code.

The input data was again normalised to a standard normal distribution and split into validation and training set by randomly choosing blocks of data as described in 3.1.2.

3.1.4. T-SNE Projection for clustering (T-SNE)

In this solution, standard T-SNE [36] was used to project all inputs (all variables except the error code) onto a 2D plane. We used the Scikit-Learn [37] implementation of T-SNE. The idea is to correlate clusters in the projection with mainly ErrorCode == 1 data points with the inputs. Experts, however, have to decide which of the variables are pure consequences of the gearbox failure and which are the causes. If one or more inputs correlate well with the given clusters this could lead to a very simple error detector.

The input data was again normalised to a standard normal distribution and split into validation and training sets by randomly choosing blocks of data as described in 3.1.2.

3.1.5. Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor (WHC-LOF)

This solution uses unsupervised methods to detect turbine failure, where the strategy is to compare simultaneously the variables of a group of wind turbines based on the SCADA data. Figure 6 shows the overview of the methodology. In the first step, we used the complete time series of bearing temperature of the five wind turbines as an input. Secondly, the data are filtered by removing the events where the bearing temperatures of the five wind turbines are not the same magnitude. The Ward Hierarchical Clustering algorithm [21] was able to identify these events by the clusters C2, C5, C6, C8, and C9, which were manually removed. Then, only the events with the same magnitude of the bearing temperature were used for training. Hence, we used the Novelty Detection with Local Outlier Factor (LOF) [22] with these training data to detect the outliers associated with the failures of the wind turbine (third step). Finally, the algorithm can identify any event where one of the turbines is different from the others as the LOF is trained with only the events with the same magnitude of the bearing temperature.

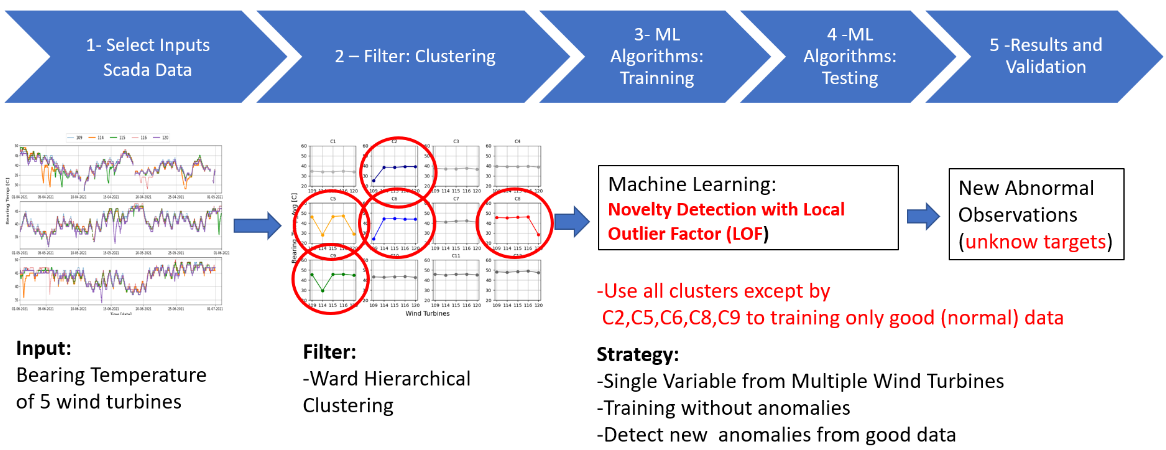

Figure 6. Overview of the WHC-LOF method that combines Ward Hierarchical Clustering (WHC) for filter process and Novelty Detection with Local Outlier Factor (LOF) for detecting the anomalies.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

3.1.6. Time-to-failure Prediction through Random Forest Regression (TTF-RFR)

This solution pre-processes data to create a new variable, named time-to-failure (TTF), which determines the time to failure of the gearbox. To this end, the data set is modified so that the problem can be treated as a supervised problem. Namely, the following steps are implemented:

1. Order all gearbox data sets with the `datetime` column in descending order.
2. Iterate through all the rows in every dataset.
3. While iterating, identify the first failure (`error code = 1`) and create the TTF variable with value 0.
4. In every subsequent row, input the difference between the previous and actual `datetime` columns in the TTF column. In this case the difference was stored in hour units.
5. Whenever the current row is an `error code`=1 row, reset the TTF value to 0.

Following this process, this solution ended up with a new TTF variable storing the remaining time until the next `error code`=1 or failure occurrence time, which can be effectively interpreted as a RUL estimation problem. For example, Figure 6 shows the TTF value for the gearbox #120.

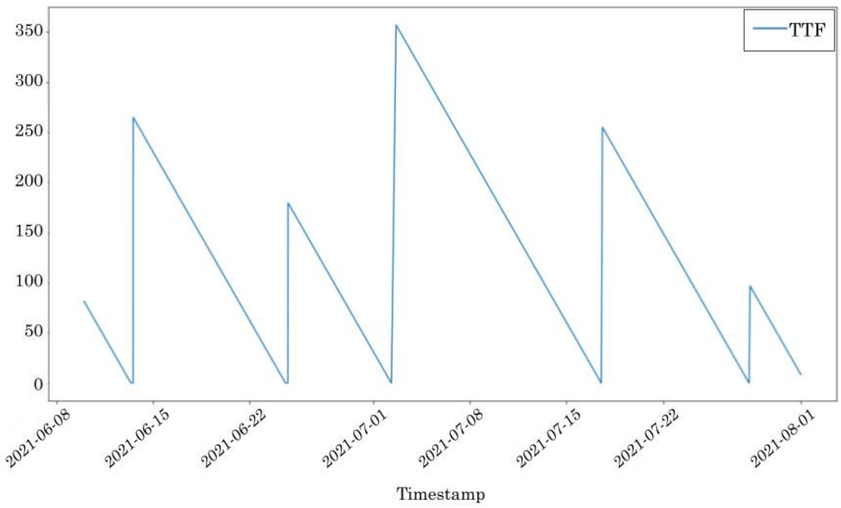

Figure 7. TTF estimation for WT #120.

The data set has been further pre-processed in order to remove (i) the data collected while the wind turbines are stopped and (ii) highly correlated variables. This has been done because, while a wind turbine is stopped, the data records that there is an error. Because of this, if these time-instants are not pre-processed and discarded, models could learn to distinguish when the wind turbine is stopped and when it is running, rather than learning to distinguish when an error occurs and when it is working correctly. In addition, variables that are highly correlated have been removed to avoid having similar or directly repeated data.

As a result, these are the selected explanatory variables that have been used to predict the TTF of the turbine: wind speed, turbulence intensity, nacelle direction, rotor speed, blade angle, generated power, power factor, ambient temperature, brake temperature, hydraulic oil temperature, bearing temperature, hydraulic pressure, gear pressure, air pressure and humidity.

After the proposed re-formulation of the problem, it is possible to resolve the TTF prediction through a supervised learning strategy. To this end, the next phase comprised training and testing the models. This was carried out following a leave-one-out cross validation methodology. In the training of every model, the data set of one gearbox is left out of the training phase and reserved for validation. After training, the data set of the gearbox left outside the training set was used to measure the testing accuracy of the trained model.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

An important detail to notice is that only the TTF values under 15 hours and above 0 hours were used for training the predictive model. Hence, the models were trained to predict the time to failure of the last 15 hours of operation, and they did not observe the data rows of the actual failures of the gearboxes. The reason for this filter was to design the regression model to (i) focus on the last hours of operation to maximize the accuracy in the most critical moments and (ii) avoid learning from data rows which contain the occurrence of actual failures and to force to predict the TTF out of normal operation data.

The model used for the prediction step was the Random Forest Regression (RFR) model, which has shown an excellent performance in various competitions, e.g. [14,38]. RFR is an ensemble of recursive trees [39]. Each tree is generated from a bootstrapped sample and a random subset of descriptors is used at the branching of each node in the tree. RFR creates a large number of trees by repeatedly resampling training data and averaging differences through voting. The RFR model has been implemented through the `sklearn` package in Python [40].

3.2. Results

3.2.1. A Convolutional Autoencoder trained with Particle Swarm Optimisation algorithm for constructing health indicators (CAE-PSO)

The proposed CAE was trained on the SCADA signals of the last 20 days leading to gearbox failure in wind turbines 114 and 120. The constructed HIs are plotted in Figure 8. Before being fed into the model, the SCADA signals were normalised to a range between 0 and 1 and the data points where the power is negative or the rotational speed is zero were dropped. The results show the ability of the proposed method to construct HIs which are monotonic ($|\tau_{MK,114}| = 0.7611$, $|\tau_{MK,120}| = 0.7623$), and, hence, correlated with the expected regime of degradation. Such HIs afford the wind farm operators the possibility to not only detect gearbox failure early, but also to estimate their RUL. However, after training the CAE with one of the two available wind turbine data with gearbox failure, the model is not able to construct a monotonic HI for the other wind turbine, i.e. it does not generalise. The reason is that training the model with only one wind turbine's data set leads to over-fitting to that specific wind turbine's operational and environmental conditions around the time of failure. Therefore, to test the performance of this model in predicting the RUL, the model needs to be trained with the data related to several wind turbines experiencing gearbox failure. This could not be done within the scope of this challenge.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

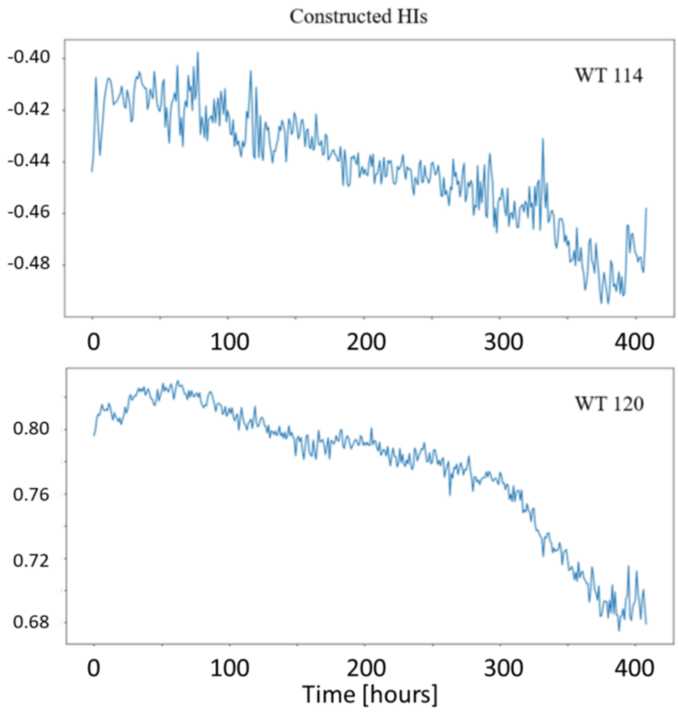

Figure 8. Constructed HIs for WT 114 and 120 - last 20 days leading to the gearbox failure.

3.2.2. Performance monitoring using Deep Neural Networks (PM-DNN)

The results of the PM-DNN solution are shown in Figure 9 (top and middle) for an example week for wind turbine number 109 (WTG 109) and WTG 120. A point is flagged as an "Outlier" if the prediction was more than 4 sigma (given by the confidence model) apart from the measurement. A "Warning" is given by a deviation of 3 sigma.

While in Figure 9-middle in some cases there is a accumulation of anomalies about 3 hours before the gearbox failure (21., 25., and 26.02.2022), one could assume that the Performance Monitoring approach could work. On the other hand there is no anomaly accumulation before the 22.02.2022, and especially the 23.02.2022. Also there is an anomaly accumulation on the 24.02.2022 with no following gearbox failure. This trend continues as can be seen in Figure 9-top. There is no significant anomaly accumulation on the 25.02.2022 and there is an accumulation on the 21.02.2022, where no gearbox failure is detected.

Table 2. Confusion matrix for WTG 109 (TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

TP 71	FN 45
FP 48	TN -

While the confusion matrix for WTG 109 is quite sobering (see Figure 2), interestingly the hydraulic oil temperature often spikes in the False Negative cases of the performance monitoring (see Figure 9-bottom). On the other hand we get a lot of oil temperature spikes without any gearbox failure. By combining this two approaches one would increase the True Positive rate but would destroy the False Positive rate.

Overall, we conclude that there are at least two kinds of gearbox failures one could maybe tracked with performance monitoring the other by monitoring the hydraulic oil temperature. Since it is not known what kind of gearbox failures exist in the data, this is hard to know for sure. So we have to conclude that there is no systematic appearance of outliers by the performance monitoring before the occurrence of gearbox failures. Therefore the model cannot be used for prediction purposes,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

because True Positive rate is too low for a reliably predictor and the combination with oil temperature monitoring would result in an unfeasible high True Negative rate.

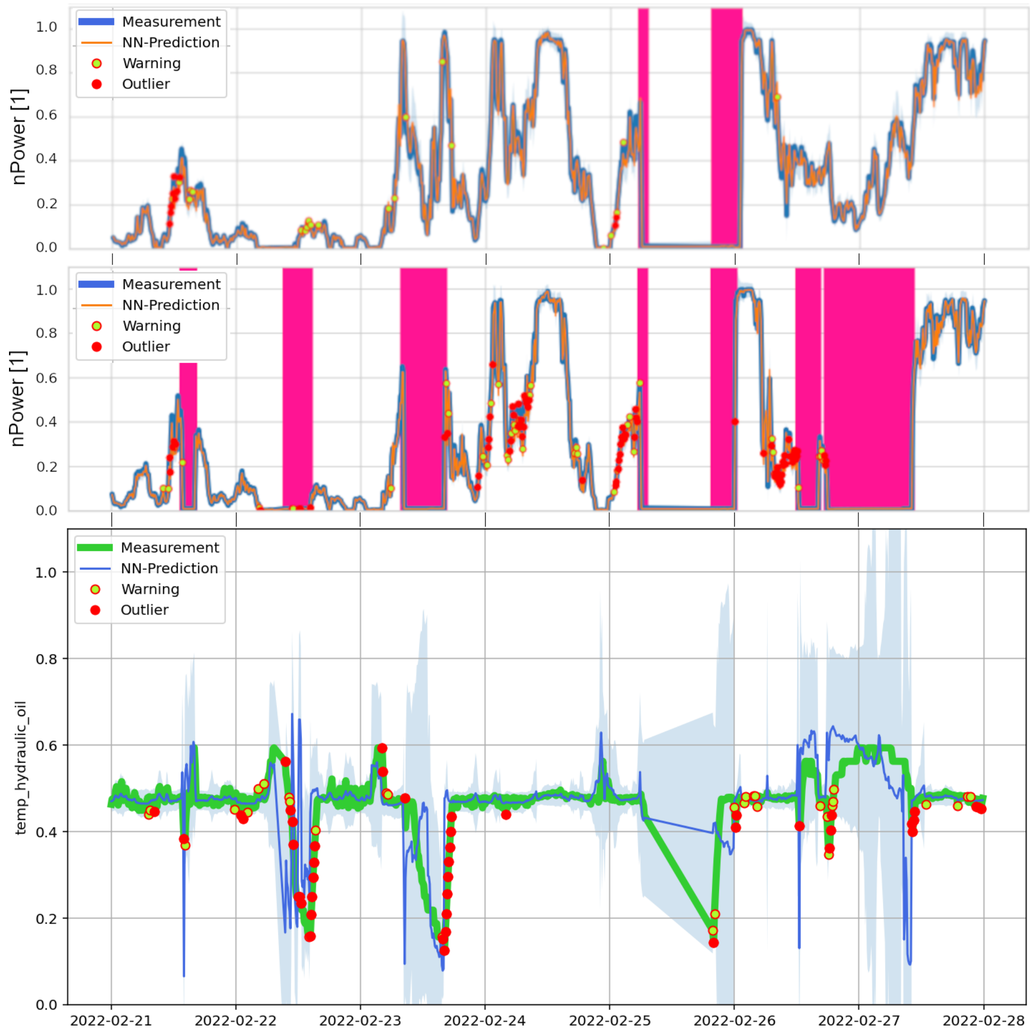

Figure 9. An example week for the WTG 120 (top) and 109 (middle) with several gearbox error codes. As can be seen for WTG 109, in some cases there is a accumulation of anomalies quite some time before the gearbox failure (21., 25., and 26.02.2022). On the other hand there is no anomaly accumulation before the 22. and especially the 23.02.2022. Also there is an anomaly accumulation at the 24.02.2022 with no following gearbox failure. Temperature of the hydraulic oil (bottom) is however spiking before the 22. The same week as for the WTG 109 is shown for the WTG 120. There is no significant anomaly accumulation on the 25.02.2022 and there is an accumulation on the 21.02.2022, where no gearbox failure is detected.

3.2.3. Direct Error Code Prediction using DNNs (DEC-DNN)

The naive direct approach worked surprisingly well, as can be seen in Figure 10. However, it had the problem that the error codes could not be predicted in advance. In fact, quite the opposite was observed. There are several examples where the error could only be predicted one time step too late (see Figure 11). The conclusion is that the gearbox failures are not predicted by precursors of the failures but by inputs that are consequences of the failure. This could perhaps be enhanced by clustering, or feature engineering, as discussed in the next solution.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

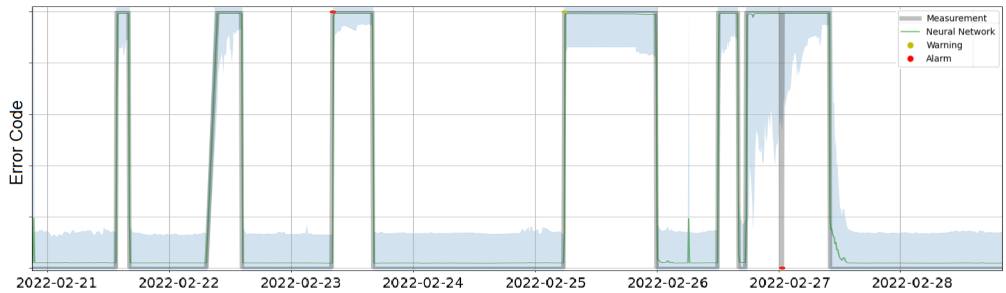

Figure 10. The direct modelling of the error codes leads to a model that can predict the gearbox failures very well as they happen but not in advance. This can be seen by the DNN (green line) following the error code (grey line) closely.

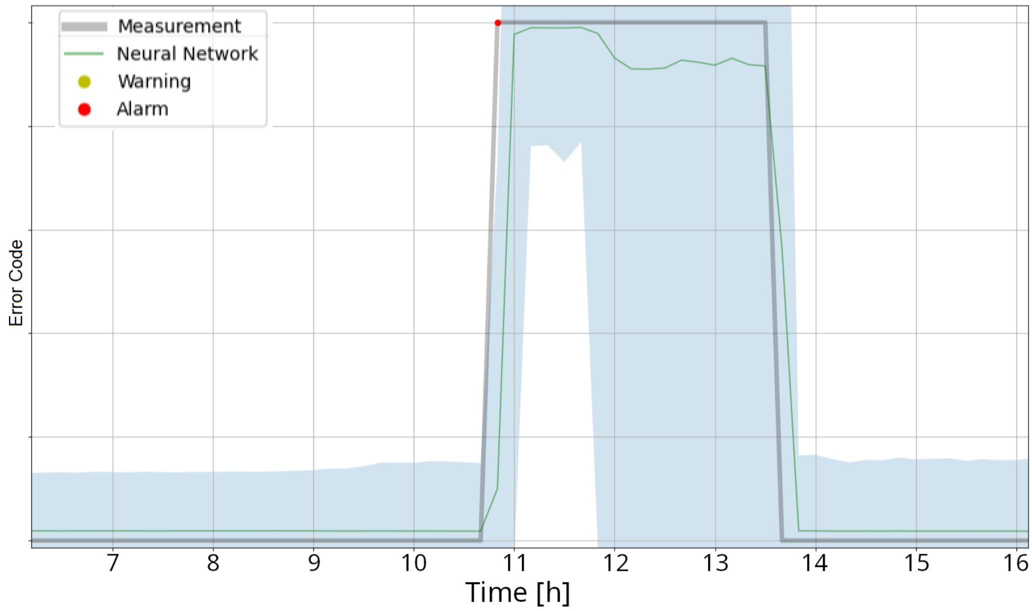

Figure 11. Sometimes the model can even predict the gearbox failure one time step too late. This can be seen in the shown instance there the error code appears 10 minutes before the DNN predicted likelihood for an error code goes up.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

3.2.4. T-SNE Projection for clustering (T-SNE)

Due to the results in Section 3.2.3, it is interesting to analyse the correlation between different inputs and the error codes, in order to investigate which of the higher correlated inputs are deemed precursors by wind energy experts. The *ErrorCode == 1* data points cluster well in a few distinct regions as shown in Figure 12. Several input variables are highly correlated with these clusters that represent high gearbox failure probability. However, the shown inputs with high correlation are all consequences of a gearbox failure, not causes. An interesting future work, in the case of the pure T-SNE projection (Section 3.1.4), would be to remove all the variables that experts think are consequences of the failure from the projection and see if the clearly structured clusters of failures are still present. This could indicate a correlation of some inputs to a gearbox failure. This would also make any connections between the remaining inputs to the error codes clearer and could provide a basis for constructing more meaningful features (Feature Engineering). This could be used as a pre-processing method for the other models presented here in the future. With the work done so far one has to assume, however, that this method is not sufficient to predict gearbox failures.

Figure 12. T-SNE projection of error codes and inputs into a 2D space. The error codes cluster well in a few distinct regions and some inputs show high correlation to these clusters. However, all inputs that were identified by this method are consequences of a gearbox failure not a cause.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

3.2.5. Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor (WHC-LOF)

Figure 13c shows the time series of temperature bearing of five wind turbines where WTG 114 (green), 109 (blue), and 120 (red) have failed, indicated by the temperature drops along the period. The WHC-LOF method was applied with the temperature bearing of those five wind turbines, and the anomalies events were considered only for WTG114. The time resolution of the alarm output is important to adjust according to the number of anomalies events accumulated by the turbine. If the time resolution is small, the WHC-LOF can indicate many False Positive (FP) alarms because of the small random anomalies associated with noises of the measurements and the normal variability of the turbine and external conditions (Figure 13a). Thus, after a sensitivity analysis of the time resolution, the WHC-LOF was configured to sum the anomalies events every week (Figure 13b), where a few weekly accumulations of anomaly events are observed before the turbine failure. They may be interpreted in two different ways. First, we have a FP alarm if compared with the same amount of accumulated error code weekly (Figure 13a) with an arbitrary threshold to trigger the alarm. Second, the significant anomalies accumulation in week bases may be considered as a prediction alarm associated with a failure event in the future with a True Positive alarm (TP). However, an expert is necessary to supervise the anomalies and indicate if it is a possible future failure of the wind turbine.

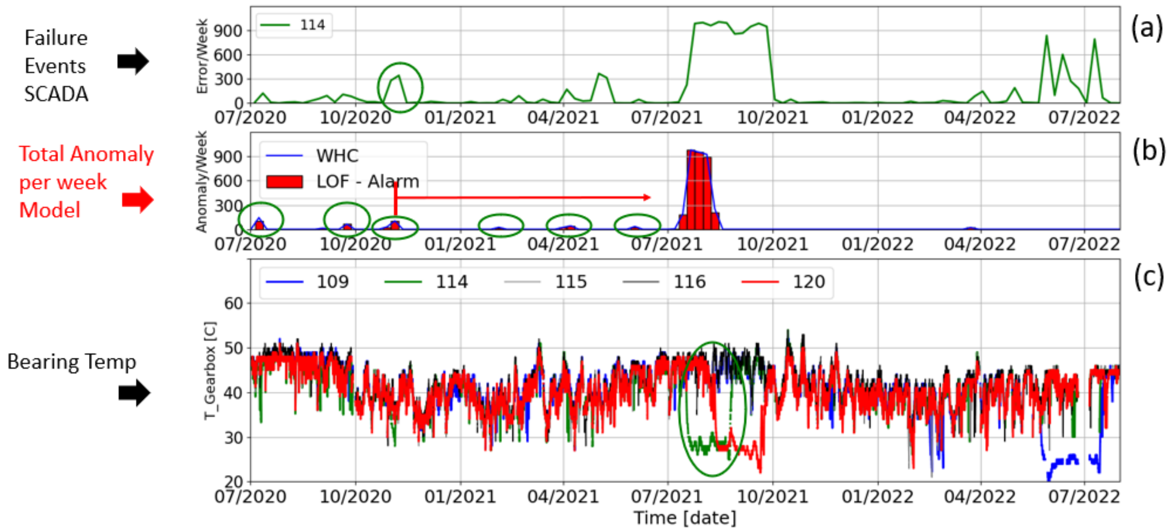

Figure 13. Error code accumulated every week by the SCADA, Failure of Turbine 114 detected by WHC-LOF method every week (b), Time series of the bearing temperature of the five wind turbines (c).

In Figure 13b, six FP alarms systematically indicate that the turbine has anomalies in its bearing temperature from July 2020 to July 2021. The possible TP alarm may be considered on 08.11.2020 with 103 anomalies/week detected by the WHC-LOF in agreement with a significant number of error codes (341) indicated by the SCADA (Figure 13a). The failure of wind turbine 114 occurred on 18.07.21, and therefore, the prediction alarm corresponds to 252 days before the turbine failure. The method positively demonstrates the prediction failure of wind turbine 114, which is necessary to understand the behaviour of the wind turbine for a long period before the turbine failure and not just the exact instant that turbine failure. The same method was applied to the EDP Wind Turbine Fault Detection Challenge [2], where the behaviour of the turbines was different from this current challenge, with only one FP detected before the turbine failure. Therefore, the complexity and variety of the behaviours before the turbine's failure demand more data sets for training the Machine Learning algorithm and for the expert to understand and learn different FPs that may be considered as TP alarms and take the right decision to prevent the turbine failure.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

3.2.6. Time-to-failure prediction using Random Forest Regression (TTF-RFR)

Model tuning was done by adjusting the number of trees in the final model (`n_estimators`), maximum depth of the tree (`max_depth`), and the number of features to consider when looking for the best split (`max_features`). These parameters were evaluated in a predefined grid of parameters. In the first iteration `n_estimators` was tested in the range [100, 250, 500, 750, 1000, 1500] and in the second iteration it was fine-tuned in the range `n_estimators`=[700, 750, 800, 850, 900] as well as `max_depth`=[3, 5, 7, 10, 15, 25] and `max_features`=[sqrt, log, auto, float, int]. The RFR hyperparameters were tuned, and the best values are obtained with `n_estimators`=850, `max_depth`=5 and `max_features`=sqrt, while leaving the remainder of hyperparameters with their default values.

Table 3 shows the mean absolute error (MAE) of the best model in every run of the cross-validation process.

Table 3. Performance statistics of the TTF prediction strategy.

Gearbox left out for testing	MAE [hours]
115	3.72
116	3.57
120	3.68
109	3.83
114	3.76

Figure 14 illustrates an example of the predictions of our model, where the y-axis is the time-to-failure in hours and x-axis is the temporal axis. The blue line is the actual TTF of the gearbox and the orange line is the prediction of the model. In some cases, such as observations 5000-5400, the model was able to predict the overall behaviour of the TTF variable, whereas in other cases, such as observations 5400-5700, it was not. Therefore, the proposed approach could be a practical TTF estimation solution, if the RFR model captures the underlying ageing process, which is not always the case.

All in all, if more information about underlying failure modes were provided, the proposed TTF-RFR approach may be a practical TTF estimate solution through capturing the ageing trajectories and their dependencies. In the current format, only the failure occurrence instants were provided, and with this information, it is challenging to accurately characterise the gearbox time-to-failure.

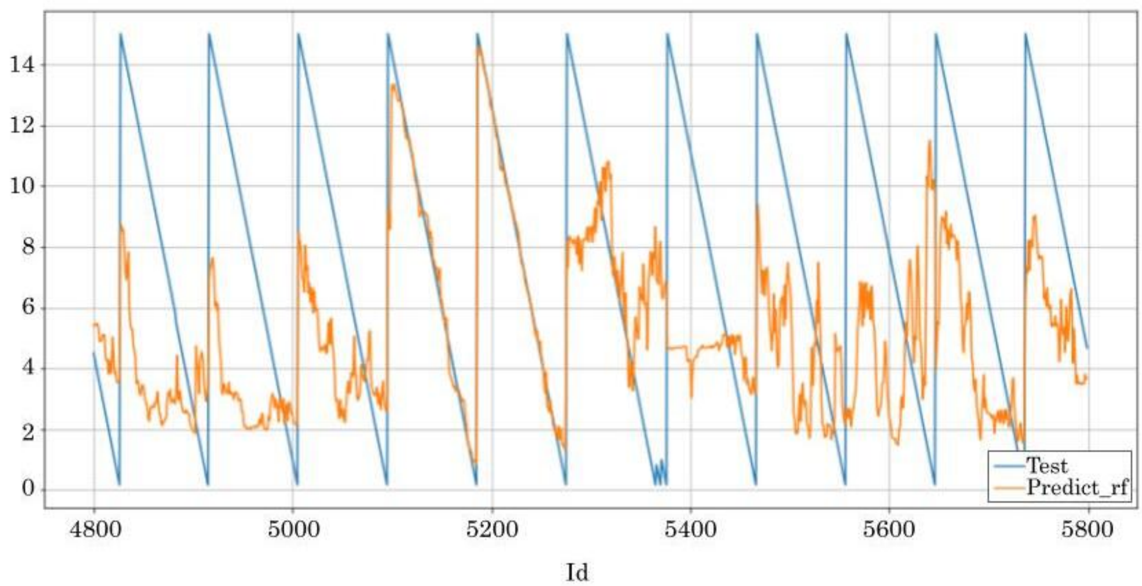

Figure 14. Difference between real and predicted TTF values for WT #109.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

3.3. Comparison and evaluation of solutions

In general, as discussed in detail in [2], the evaluation of time-series anomaly detection algorithms is a difficult task. This is due to the sensitivity of the algorithm performance to the alarm threshold [41] as well as because false and missed alarms can have very different implications. For SCADA data based wind turbine fault detection, the most common measures include root mean squared error (RMSE) [42], mean absolute percentage error (MAPE) [43] and mean absolute error (MAE). Classification models are usually evaluated using accuracy, sensitivity, specificity and F1 [44–46]. However, these measures do not take into account the costs associated with inspections, repairs and replacements. As this step is very specific to the asset owner, there is no agreed-upon method for doing this in the literature. In the EDP Challenge, a simple method for estimating costs of True positives (TP), False negatives (FN) and False positives (FP) for different wind turbine components was applied [2].

In this work, the goal was to "train, test and validate methods that will provide clear indicators of an upcoming gearbox-related fault, as well as/or a horizon-based probability of the event occurring". WinJi wanted to "push the performance of such predictions to at least a monthly horizon". However, the submitted solutions did not achieve this directly. A summary of the main results and learning outcomes related to each solution is shown in Table 4.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

Table 4. Summary of main results and learning outcomes.

Solution	Summary of main results	Learning outcomes related to data sharing
CAE-PSO	HIs correlate with degradation, therefore can be used for early failure detection and RUL estimation. However, the result was not generalisable to other wind turbines. The model needs to be trained with data of several wind turbines.	More data sets are required from multiple WTGs experiencing gearbox failure.
PM-DNN	No systematic appearance of outliers before the occurrence of gearbox failures. The model cannot be used for prediction purposes, because True Positive rate is too low for a reliably predictor and the combination with oil temperature monitoring would result in an unfeasibly high True Negative rate.	More information about the optimisation goals (i.e. costs of TPs, FPs and FNs) are needed. More details of failure types are required.
DEC-DNN	The gearbox failures were not predicted by precursors of the failures but by inputs that are consequences of the failure.	None.
T-SNE	Several input variables were highly correlated with these clusters that represent high gearbox failure probability. The shown inputs with high correlation are all consequences of a gearbox failure, not causes. The method is therefore not sufficient to predict gearbox failures.	More details of failure types and detailed analysis with experts needed.
WHC-LOF	The method has potential but the complexity and variety of the behaviours before the turbine's failure demand more data sets for training the algorithm and for the expert to understand and learn different FPs that may be considered as TP alarms and take the right decision to prevent the turbine failure.	More details of failure types and detailed analysis with experts needed.
TTF-RFR	With the information provided here, it was challenging to accurately characterise the gearbox time-to-failure. Time to failure could be predicted with MAE of 3.57-3.72 h. If more information about underlying failure modes were provided, the approach may be a practical solution through capturing the ageing trajectories and their dependencies.	More details of failure types and detailed analysis with experts needed.

In general, it can be concluded that the CAE-PSO, the PM-DNN, the WHC-LOF and the TTF-RFR solutions appear to have a high potential to reach the goals of the challenge. However, there are a number of concrete things that would need to have been done by the Challenge Providers:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

- Access to more details of the different failure types. This could be done in future WeDoWind Challenges; however, a trade-off needs to be made by Challenge Providers between anonymising the data and providing enough information to allow the challenge to be completed.
- Access to multiple data sets from more WTGs experiencing gearbox failure. Again, this could be done in future WeDoWind Challenges, but in this case a decision balancing the amount of provided data and the expected quality of the results needs to be made by the Challenge Provider.
- Provision of a model or rule relating fault detection times or remaining useful lifetime to the estimated costs for repairs, replacements and inspections similar to the EDP one mentioned in Section 3.3. This could be done by the Challenge Provider in the future. Alternatively, this could even be developed by Solution Providers at the start of a Challenge, and facilitated by the moderator.
- Provision of a clear strategy for training and test periods in advance, as is common on other challenge-based platforms such as Kaggle (and as was the case in the previous WeDoWind EDP Challenge), by providing separate training and test data, or by defining how the participants should choose training and test data. In the future, if this is not provided by the Challenge Provider, it could be decided by the participants, and facilitated by the moderator.
- Participation in the Challenge by the Challenge Providers, and provision of details of their results and a description of their method. This can be recommended by the moderators at the start of a WeDoWind Challenge in the future. If there are confidentiality issues, they can take part in a reduced way, i.e. by only providing limited results.
- Provision of a pre-defined template or requirements, and even a pre-written analysis code, in order to improve the comparison process. This can be provided by Challenge Providers in future WeDoWind Challenges.

Additionally, the WeDoWind Challenge moderator could have supported this process more actively, as well as focused more on keeping the exact goal of the challenge in sight during the Challenge running.

It can be concluded that there are a number of issues and difficulties that come up when an attempt is made to share data for the purpose of comparing and evaluating different SCADA data based wind turbine fault detection models, such as in this WeDoWind Challenge study. This work has allowed a set of best practice data sharing guidelines to be developed, as described below in Section 4.

Regarding the WeDoWind Method, we have shown that it can be applied to successfully bring together a pool of talented and diverse people with different backgrounds to tackle a problem together. It exposes them to the realities of having to deal with unstructured and uncharacterised data, rather than nicely-prepared simulation data. While further work is required to use the WeDoWind Method to directly produce easily-applicable open-source code, as discussed above, it nevertheless provides Challenge Providers with new insights and ideas for the improvement of their models. Further case studies with a wide range of partners are underway. Ultimately, this allows the three "Grand Challenges of digitalisation in wind energy" to be addressed in a holistic manner. The results will be used to iteratively enhance and improve the best practice data sharing guidelines.

4. Best practice data sharing guidelines for wind turbine fault detection model evaluation

The best practice data sharing guidelines have been developed as a list for practitioners to use directly, as summarised in Table 5.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

Table 5. Best practice data sharing guidelines for wind turbine fault detection model evaluation.

Number	Guideline	Recommendation	Comments
1	Some information can be anonymised or normalised to maintain confidentiality	Wind turbine type, location and rated power are recommended	Reducing faults to binary codes can be very limiting for participants, especially if remaining useful lifetime is required
2	The details provided about failure types should be sufficient enough to allow models to be trained, in order to get the best compromise between information and confidentiality	3-4 fault types are recommended	The "Basic concepts and taxonomy of dependable and secure computing" published by IEEE [47] is useful for describing failure types
3	Sufficient data sets containing multiple failures of various wind turbines should be provided, in order to get the best compromise between information and confidentiality	At least 3-4 different faults should occur in 3-4 different wind turbines	-
4	A model or rule relating fault detection times or remaining useful lifetime to the estimated costs for repairs, replacements and inspections should be provided in advance.	A cost model for repairs, replacements and inspections similar to the EDP one is recommended [2].	If this cannot be done, the participants should make sure that this is discussed and agreed on at the start of the challenge.
5	A clear strategy for training and test periods should be provided in advance.	The provided data can be split into training and test data sets in advance.	If this cannot be done, the participants should make sure that this is discussed and agreed on at the start of the challenge.
6	The challenge provider should take part in the challenge.	In the best case, they should share both results and a description of their method.	If there are confidentiality issues, they can take part in a reduced way, i.e. by only providing results.
7	The challenge moderator should focus more on the goal of the challenge and actively do things to help the participants reach the goal.	Requiring the results to be submitted in a certain format or providing a template.	An analysis code can also be provided.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

5. Conclusions

In this work, the challenges related to comparing and evaluating different SCADA data based wind turbine fault detection models were investigated by carrying out a new case study, the "WinJi Gearbox Fault Detection Challenge", based on the WeDoWind Method. The WeDoWind Method has previously been shown to create tangible incentives to motivate and empower different types of people from all over the world to actually share data and knowledge in practice.

The goal of the challenge was to help WinJi, the Challenge Providers, to push the performance of their predictions to at least a monthly horizon, by making use of the provided SCADA data in order to train, test and validate methods that will provide clear indicators of an upcoming gearbox related fault, as well as/or a horizon-based probability of the event occurring. Six new solutions were submitted to the challenge, using approaches including a Convolutional Autoencoder trained with Particle Swarm Optimisation algorithm for constructing health indicators (CAE-PSO), performance monitoring using Deep Neural Networks (PM-DNN), Direct Error Code Prediction using DNNs (DEC-DNN), T-SNE Projection for clustering (T-SNE), Combined Ward Hierarchical Clustering and Novelty Detection with Local Outlier Factor (WHC-LOC) and Time-to-failure prediction using Random Forest Regression (TTF-RFR).

A comparison and evaluation of the results showed that, in general, some of the approaches (CAE-PSO, PM-DNN, WHC-LOF and TTF-RFR) appear to have a high potential to reach the goals of the challenge. However, there were a number of concrete things that would need to have been done by the challenge providers and the challenge moderators in order to ensure success. This includes enabling access to more details of the different failure types, access to multiple data sets from more WTGs experiencing gearbox failure, provision of a model or rule relating fault detection times or a remaining useful lifetime to the estimated costs for repairs, replacements and inspections, provision of a clear strategy for training and test periods in advance, participation in the challenge and provision of details of their results and description of their method as well as provision of a pre-defined template or requirements.

This work presents some of the difficulties encountered when enabling cooperation between organisations and connecting people and data for a successful implementation of digitalisation in wind energy. Thus, these learning outcomes were used directly to define a set of best practice data sharing guidelines for wind turbine fault detection model evaluation, in order to improve model evaluation and data sharing in the future.

Author Contributions: Conceptualization, S.B., C.H.; methodology, U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S.; software, U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S.; validation, U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S.; formal analysis, U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S.; investigation, S.B., U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S.; resources, S.B., U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S., C.H.; data curation, U.I, O.S., J.O., E.Z., J.I.A., A.E.M., F.S., Y.S.; writing—original draft preparation, S.B., J.I.A., A.E.M., F.S., Y.S.; writing—review and editing, S.B., J.I.A., A.E.M., F.S., Y.S., C.H.; visualization, S.B., J.I.A., A.E.M., F.S., Y.S.; supervision, S.B., C.H.; project administration, S.B.. All authors have read and agreed to the published version of the manuscript.

Funding: J. I. A. and E. S. are partially supported by the Basque Government (KK-2022-00106) and J. I. A. is funded by Juan de la Cierva Incorporacion Fellowship, Spanish State Research Agency (grant No. IJC2019-039183-I). The other authors received no external funding.

Data Availability Statement: The code developed in this work is available under reference [48]

Acknowledgments: We acknowledge the kind support and internal funding of the Eastern Switzerland University of Applied Sciences for this work.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

References

1. Clifton, A.; Barber, S.; Bray, A.; Enevoldsen, P.; Fields, J.; Sempreviva, A.M.; Williams, L.; Quick, J.; Purdue, M.; Totaro, P.; Ding, Y. Grand Challenges in the Digitalisation of Wind Energy. *Wind Energy Science* **2022**, *In Review*. doi:wes-2022-29.
2. Barber, S. Co-Innovation for a Successful Digital Transformation in Wind Energy Using a New Digital Ecosystem and a Fault Detection Case Study. *Energies* **2022**, *xx*. doi:xxx.
3. Maria, S.A.; Allan, V.; Christian, B.; Robert, V.D.; Gregor, G.; Kjartansson, D.H.; Pilgaard, M.L.; Mattias, A.; Nikola, V.; Stephan, B.; Javier, S.R.; Pawel, G.; Inge, R.T.; Christian, B.H.; Willem, W.J.; W., H.K. Taxonomy and metadata for wind energy Research &Development, 2017. Integrated Research Programme on Wind Energy, IRPWIND Grant agreement no 609795, , collaborative project, Start date: 01st December 2013 Work Package 2 - Deliverable D2.3 Lead Beneficiary: DTU Delivery date: 12 December 2017 Dissemination level: PU The research leading to these results has received funding from the European Union Seventh Framework Programme under the agreement 609795., doi:10.5281/zenodo.1199489.
4. Barber, S.; Clark, T.; Day, J.; Totaro, P. The IEA Wind Task 43 Metadata Challenge: A roadmap to enable commonality in wind energy data, 2022. doi:10.5281/zenodo.6457039.
5. Bresciani, S.; Ciampi, F.; Meli, F.; Ferraris, A. Using big data for co-innovation processes: Mapping the field of data-driven innovation, proposing theoretical developments and providing a research agenda. *International Journal of Information Management* **2021**, *60*, 102347. doi:https://doi.org/10.1016/j.ijinfomgt.2021.102347.
6. Lee, S.M.; Olson, D.L.; Trimi, S. Co-innovation: convergenomics, collaboration, and co-creation for organizational values. *Management decision* **2012**.
7. Dao, C.; Kazemtabrizi, B.; Crabtree, C. Wind turbine reliability data review and impacts on levelised cost of energy. *Wind Energy* **2019**, *22*, 1848–1871, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.2404]. doi:10.1002/we.2404.
8. Zaher, A.; McArthur, S.; Infield, D.; Patel, Y. Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology* **2009**, *12*, 574–593.
9. Butler, S.; Ringwood, J.; O'Connor, F. Exploiting SCADA system data for wind turbine performance monitoring. 2013 Conference on Control and Fault-Tolerant Systems (SysTol). IEEE, 2013, pp. 389–394.
10. Kusiak, A.; Verma, A. Analyzing bearing faults in wind turbines: A data-mining approach. *Renewable Energy* **2012**, *48*, 110–116.
11. Sun, P.; Li, J.; Wang, C.; Lei, X. A generalized model for wind turbine anomaly identification based on SCADA data. *Applied Energy* **2016**, *168*, 550–567.
12. Bangalore, P.; Letzgus, S.; Karlsson, D.; Patriksson, M. An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox. *Wind Energy* **2017**, *20*, 1421–1438.
13. Bach-Andersen, M.; Rømer-Odgaard, B.; Winther, O. Flexible non-linear predictive models for large-scale wind turbine diagnostics. *Wind Energy* **2017**, *20*, 753–764.
14. Liu, Y.; Chen, H.; Zhang, L.; Feng, Z. Enhancing building energy efficiency using a random forest model: A hybrid prediction approach. *Energy Reports* **2021**, *7*, 5003–5012. doi:https://doi.org/10.1016/j.egyr.2021.07.135.
15. Stetco, A.; Dinmohammadi, F.; Zhao, X.; Robu, V.; Flynn, D.; Barnes, M.; Keane, J.; Nenadic, G. Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy* **2019**, *133*, 620–635. doi:https://doi.org/10.1016/j.renene.2018.10.047.
16. Schlechtingen, M.; Santos, I.F. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical systems and signal processing* **2011**, *25*, 1849–1875.
17. Fu, J.; Chu, J.; Guo, P.; Chen, Z. Condition Monitoring of Wind Turbine Gearbox Bearing Based on Deep Learning Model. *IEEE Access* **2019**, *7*, 57078–57087. doi:10.1109/ACCESS.2019.2912621.
18. Gougam, F.; Rahmoune, C.; Benazzouz, D.; Varnier, C.A.C.; Nicod, J.M. Health Monitoring Approach of Bearing : Application of Adaptive Neuro Fuzzy Inference System (ANFIS) for RUL-Estimation and Autogram Analysis for Fault-Localization. *2020 Prognostics and Health Management Conference (PHM-Besançon)* **2020**, pp. 200–206.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

19. Li, X.; Teng, W.; Peng, D.; Ma, T.; Wu, X.; Liu, Y. Feature fusion model based health indicator construction and self-constraint state-space estimator for remaining useful life prediction of bearings in wind turbines. *Reliability Engineering & System Safety* **2023**, *233*, 109124. doi:https://doi.org/10.1016/j.ress.2023.109124.
20. Xu, Z.; Bashir, M.; Liu, Q.; Miao, Z.; Wang, X.; Wang, J.; Ekere, N. A novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model. *Computers & Industrial Engineering* **2023**, *176*, 108999. doi:https://doi.org/10.1016/j.cie.2023.108999.
21. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **1963**, *58*, 236–244.
22. Breunig, M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. Proceedings Of The 2000 Acm Sigmod International Conference On Management Of Data. ACM, 2000, pp. 93–104.
23. Ahmed, I.; Dagnino, A.; Ding, Y. Unsupervised Anomaly Detection Based on Minimum Spanning Tree Approximated Distance Measures and its Application to Hydropower Turbines. *IEEE Transactions on Automation Science and Engineering* **2019**, *16*, 654–667. doi:10.1109/TASE.2018.2848198.
24. Dao, P.B. A CUSUM-Based Approach for Condition Monitoring and Fault Diagnosis of Wind Turbines. *Energies* **2021**, *14*. doi:10.3390/en14113236.
25. Xu, Q.; xiang Lu, S.; Zhai, Z.; Jiang, C. Adaptive fault detection in wind turbine via RF and CUSUM. *Iet Renewable Power Generation* **2020**, *14*, 1789–1796.
26. Pandit, R.; Astolfi, D.; Hong, J.; Infield, D.; Santos, M. SCADA data for wind turbine data-driven condition/performance monitoring: A review on state-of-art, challenges and future trends. *Wind Engineering* **0**, *0*, 0309524X221124031, [https://doi.org/10.1177/0309524X221124031]. doi:10.1177/0309524X221124031.
27. Astolfi, D.; Pandit, R.; Terzi, L.; Lombardi, A. Discussion of Wind Turbine Performance Based on SCADA Data and Multiple Test Case Analysis. *Energies* **2022**, *15*. doi:10.3390/en15155343.
28. Maron, J.; Anagnostos, D.; Brodbeck, B.; Meyer, A. Artificial intelligence-based condition monitoring and predictive maintenance framework for wind turbines. *Journal of Physics: Conference Series* **2022**, *2151*, 012007. doi:10.1088/1742-6596/2151/1/012007.
29. Pohlert, T. Non-Parametric Trend Tests and Change-Point Detection, 2015. doi:10.13140/RG.2.1.2633.4243.
30. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *CoRR* **2014**, *abs/1404.7828*, [1404.7828].
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
32. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc., 2019; pp. 8024–8035.
33. Sehnke, F.; Osendorfer, C.; Rückstieß, T.; Graves, A.; Peters, J.; Schmidhuber, J. Parameter-exploring policy gradients. *Neural Networks* **2010**, *23*, 551–559.
34. Sehnke, F. Efficient baseline-free sampling in parameter exploring policy gradients: Super symmetric pgpe. International Conference on Artificial Neural Networks. Springer, 2013, pp. 130–137.
35. Sehnke, F.; Zhao, T. Baseline-free sampling in parameter exploring policy gradients: Super symmetric pgpe. In *Artificial Neural Networks*; Springer, 2015; pp. 271–293.
36. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
38. Nagy, G.I.; Barta, G.; Kazi, S.; Borbély, G.; Simon, G. GEFCom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach. *International Journal of Forecasting* **2016**, *32*, 1087–1093. doi:https://doi.org/10.1016/j.ijforecast.2015.11.013.
39. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
41. Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T.G.; Müller, K.R. A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE* **2021**, *109*, 756–795. doi:10.1109/JPROC.2021.3052449.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2023 doi:10.20944/preprints202303.0239.v1

42. Orozco, R.; Sheng, S.; Phillips, C. Diagnostic Models for Wind Turbine Gearbox Components Using SCADA Time Series Data. 2018 IEEE International Conference on Prognostics and Health Management (ICPHM), 2018, pp. 1–9. doi:10.1109/ICPHM.2018.8448545.
43. Wang, L.; Zhang, Z.; Long, H.; Xu, J.; Liu, R. Wind Turbine Gearbox Failure Identification With Deep Neural Networks. *IEEE Transactions on Industrial Informatics* **2017**, *13*, 1360–1368. doi:10.1109/TII.2016.2607179.
44. Stetco, A.; Dinmohammadi, F.; Zhao, X.; Robu, V.; Flynn, D.; Barnes, M.; Keane, J.; Nenadic, G. Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy* **2019**, *133*, 620–635. doi:https://doi.org/10.1016/j.renene.2018.10.047.
45. Arcos Jiménez, A.; Gómez Muñoz, C.Q.; García Márquez, F.P. Machine Learning for Wind Turbine Blades Maintenance Management. *Energies* **2018**, *11*. doi:10.3390/en11010013.
46. Tang, B.; Song, T.; Li, F.; Deng, L. Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine. *Renewable Energy* **2014**, *62*, 1–9. doi:https://doi.org/10.1016/j.renene.2013.06.025.
47. Avizienis, A.; Laprie, J.C.; Randell, B.; Landwehr, C. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing* **2004**, *1*, 11–33. doi:10.1109/TDSC.2004.2.
48. Barber, S. GitLab repository "WeDoWind - WinJi Gearbox Failure Detection", 2023.

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.