20th International Conference on Renewable Energies and Power Quality (ICREPQ'22)
Vigo (Spain), 27th to 29th July 2022
Renewable Energy and Power Quality Journal (RE&PQJ)
ISSN 2172-038 X, Volume No.20, September 2022

Main Bearing Fault Prognosis in Wind Turbines based on Gated Recurrent Unit Neural Networks

Á. Encalada-Dávila[1], C. Tutivén[1], Luis Moyón[2], B. Puruncajas[1] and Y. Vidal[3,4]

[1] Mechatronics Engineering
Faculty of Mechanical Engineering and Production Science, FIMCP
Escuela Superior Politécnica del Litoral, ESPOL
Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
Phone number: +593 9 9103 5259, e-mail: **{angaenca, cjtutive, bpurunca}@espol.edu.ec**

[2] Universidad ECOTEC, Km. 13.5 Samborondón, Samborondón, EC092303, Ecuador
Phone number: +593 04 3723400, e-mail: **amoyon@hotmail.com**

[3] Control, Modeling, Identification and Applications, CoDAlab
Department of Mathematics, Escola d'Enginyeria de Barcelona Est, EEBE
Universitat Politècnica de Catalunya, UPC
Campus Diagonal-Besós (CDB) 08019, Barcelona, Spain

[4] Institut de Matemàtiques de la UPC - BarcelonaTech, IMTech
Pau Gargallo 14, 08028 Barcelona, Spain
Phone number: +34 934 137 309, e-mail: **yolanda.vidal@upc.edu**

Abstract. The transition from onshore to offshore wind farms is an imminent fact in the future. It supposes to face hard challenges like difficulties to carry out offshore maintenance operations due to increased downtime (because of several causes like continuously bad environmental conditions) on wind farms. That is why, there is a need to improve maintenance and monitoring practices like those involved in condition-based area. This work proposes a methodology based on three key points: (i) a semi-supervised model built from a gated recurrent unit (GRU) neural network and by using only healthy real SCADA data, (ii) propose a fault prognosis indicator (FPI) to trigger warnings or fault alarms as such, and (iii) detect the main bearing fault several months in advance on a faulty wind turbine. The reported results show the excellent performance of the GRU trained model to predict the main bearing temperature as output by exploiting the capabilities of GRUs (recurrent-based neural networks) to decide what information to forget or preserve through time. In the FPI construction, the use of exponentially weighted moving average (EWMA) helps at the results to avoid the presence of false alarms that is very useful in any detection strategy. Finally, the stated methodology lets to detect the main bearing fault on a WT two months in advance at least, which contributes to plan maintenance actions ahead of time. Furthermore, in this way, the lifespan of this large component may be extended and wind turbine's uptime may increase in a significant percentage.

Key words

Wind turbine, fault prognosis, main bearing, SCADA data, GRU neural networks.

1. Introduction

Currently, the leaders in energy production from wind energy are China and USA. Along the last years, installed capacity of wind energy has set two main sources: onshore and offshore. The International Energy Agency's (IEA) Sustainable Development Scenario [1] forecasts that the installed capacity in offshore wind generation would be increased from 19 GW (in 2018) to 127 GW in 2040. It is clearly expected that with an increase in wind energy generation, the scale of operation and maintenance (O&M) costs will increase also. Likewise, the transition from onshore wind farms (which currently dominate the wind energy market) to offshore wind farms will bring great challenges like facing increased downtime and difficulties to carry maintenance operations out.

Just to keep in mind, the National Renewable Energy Laboratory (NREL) states that OM costs for USA's offshore wind energy are around 83-250 USD/kW/year [2]. On average, offshore wind turbines (WTs) have a failure rate of 10 failures per WT per year, where 17.5 % corresponds to major repairs and 2.5 % are due to major replacements [3]. Furthermore, one of the WT larger components that faults is the main bearing, which supports the low-speed shaft. Thus, in short, this paragraph supposes an important need to be neater and more proactive in WT maintenance practices to decrease O&M costs and make this market more sustainable through time.

Typically, WT maintenance practices are split in corrective, scheduled, and condition-based. This last one is not very common, but it has become more popular in last years, which involves actively monitoring WT components and attempts to forecast or detect failures in advance. Thus, it is noteworthy that once failures are predicted, WT maintenance actions can

be scheduled ahead of time [4].

Following those ideas, this work proposes a methodology to early detect the main bearing fault in WTs. The methodology is based on the use of only real SCADA data and a model built on a gated recurrent unit (GRU) neural network. Thus, the following paragraph intends to briefly address several studies in the field which made a baseline to develop this work.

In [5] by using support vector machine (SVM) classifiers the main bearing fault is diagnosed, or as in [6] where simultaneous multiple faults are also detected by using SVM. However, the deployment of trained models with supervised algorithms may be very dough since in real world the construction of historical labeled-data is challenging and a bit risky. On the other hand, the use of SCADA data is supported by several works like [5], [7]. Now, delving into deep learning, some studies have been carried out, for instance, by using artificial neural networks (ANNs) to detect the main bearing fault ahead of time [8].

To conclude, this work approaches some key points which are mentioned as follows:

- The training of a semi-supervised model that is based on only healthy data, establishing a normality model. It avoids the likelihood of imbalance data problems when supervised algorithms are used.
- A robust trained model that precludes the presence of false alarms in the main bearing fault detection process.
- A sturdy fault prognosis methodology that treats the presence of false alarms and the setup of thresholds which let to get a warning or fault alarm depending on the severity of the main bearing fault through time.
- The methodology predicts the main bearing fault two months in advance at least, that lets WT operators plan maintenance actions ahead of time.

The rest of the paper is organized as follows: Section 2 where an overview on the installed wind farm is stated as well as a brief explanation about main bearing faults is given. Next, Section 3 describes all about collected and used SCADA data in this study. Likewise, the data preprocessing is explained there. Section 4 introduces the use of a GRU neural network to build the semi-supervised model. Then, the computation of the fault prognosis indicator (FPI) is given also in this section. Section 5 contains the results and its respective discussion. Finally, the conclusions of this study are given in Section 6.

2. Wind Farm Overview and Main Bearing Fault

The wind farm is composed by 18 WTs and each one is characterized by a diameter of 101 m and a nominal power of 2.3 MW. In Figure 1 can be seen the principal components of a WT. Among the characteristics of this kind of WTs, they have 3 blades, a cut-in wind speed of 3 m/s, a cut-off wind speed of 12 m/s, and a rated wind speed of 12 m/s. For purpose of this study, the component to be highlighted is the main bearing, which is a double-spherical roller type. This

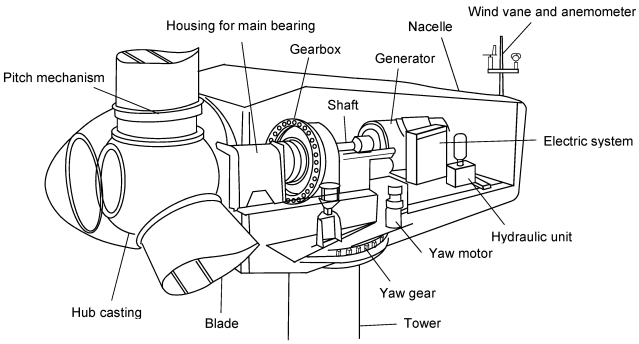

Fig. 1: Main components of a WT. Adapted from [9].

Table I. Bearing failure modes based on the ISO 15243 standard.

Failure mode	Classification
Fatigue	Subsurface-initiated fatigue
	Surface-initiated fatigue
Wear	Abrasive wear
	Adhesive wear
Corrosion	Moisture corrosion
	Fretting corrosion
	False brinelling
Electrical erosion	Excessive current erosion
	Current leakage erosion
Plastic deformation	Overload
	Indentations
Fracture and cracking	Forced fracture
	Fatigue fracture
	Thermal cracking

sort of bearings is suitable for applications where high loads and very low speeds occur. Furthermore, these bearings are designed to withstand loads with variable direction and low friction that involves a longer lifespan and minor energy loss.

Elements of machines like shafts need to rotate and require additional components to comply to this function, and so the bearings take relevance. Bearings normally are integrated by four elements: a cage, an inner race, an outer race, and a rolling. Evidently, these components are always under mechanical stress by different forces like frictional, impact, inertial, or centrifugal. Specifically on a WT, the main bearing supports the main shaft, both are large components. The Swedish bearing and seal manufacturing company called SKF has classified the different bearing failure modes based on the ISO 15243 standard. Table I shows a summary of the failure modes for this component. Additionally, in [8] it is explained with detail each bearing failure mode including some illustrative pictures.

In a nutshell, all bearing failures present a starting point from which a series of anomalous behavior begins. This behavior can be monitored looking at changes in main bearing or shaft temperatures, output power, vibrations, pressure, etc. A notable symptom may be the heat release, that is why this study aims to predict the main bearing fault some months in advance. This early alert may help to give suitable maintenance and avoid that the component seriously fails in the future.

3. Real SCADA Data Preprocessing

For the scope of the study, two WTs were selected, one is healthy and the other is faulty (i.e., there was a main bearing fault occurred on June 11, 2018). The data gathered samples (rated each 10 min) from January 1 in 2015 up to December 31 in 2018. As it is known, SCADA data normally contain measurements associated to different sections like environmental, electrical, temperature, hydraulic, and control. The data are naturally acquired with a frequency of 1 Hz, however, these are averaged and stored with a period time of 10 min, so the mean, maximum, minimum, and standard deviation for most of the variables are computed and registered.

The basis of the study is using only the variables with their mean values. Furthermore, as the component under analysis is the main bearing, the closest variables to this element were taken into account to join an exogenous variable corresponding to the wind speed. Next, the selected variables are shown:

- wtc_MainBTmp_mean: the mean main bearing temperature, in °C.
- wtc_GenBeTm_mean: the mean generator bearing temperature, in °C.
- wtc_GeOilTmp_mean: the mean gearbox oil temperature, in °C.
- wtc_PrWindSp_mean: the mean primary wind speed, in m/s.

Besides the SCADA data, information on maintenance and repair actions is essential since it lets to locate the failure type, start and end dates for the work orders, affected subsystems, and actions performed. To summarize, this information characterizes if a WT is healthy or not, i.e., if any records about the failure are registered or not.

A. Deseasonalizing and Data Imputation

As above-mentioned, three of the selected variables are temperatures. The data run throughout the years and the seasonality is present since if in the middle of a year the weather is hot and in the another middle the weather is cold the temperatures are not obviously the same and this behavior is replicated each year. Then, the ambient temperature is subtracted from all variables related to temperature to avoid this problem [10].

Real data is normally noisy, thus, data imputation techniques are needed to face this problem. The more basic strategies like filling missing values with the mean, median, or mode are not suitable in this case since they introduce a bias in the data. More advanced techniques are recommended since they offer a better treatment of missing values based on data surround-trends. Thus, the Piecewise Cubic Hermite Interpolating Polynomial is proposed [8], which guarantees to maintain the data shape and its monotonicity. As at the beginning or final of the dataset there may be NaN values, these are filled with their closest value.

B. Data Split

To success in the construction of a machine learning model, data split is the basis. Data are normally split in three subsets: training, validation, and testing. In this case, only the healthy data is employed to train and validate the model since the goal is building a normality model. Likewise, regarding to the seasonality problem, the training and validation datasets must cover at least one full year. With the trained model, the testing dataset (it contains data of the main bearing fault) is employed instead to examine the performance of the model.

To be detailed, the SCADA data are split as follows:

- Training: from January 2015 to September 2017 (144576 samples).
- Validation: from October 2017 to December 2017 (13248 samples).
- Testing: full year 2018 (52560 samples).

C. Data Normalization

As the data integrate many different variables, as these come from different sources, their orders of magnitude vary. On a machine learning model in general it is suitable that all variables work under the same scale since it avoids some variables add more weight and cause imbalance. In this manuscript, the data are normalized employing the min-max strategy that scales the values in a range from 0 to 1, without losing proportionality [11].

4. Fault Detection Methodology

This section addresses two key parts: the construction of the normality model based on a GRU neural network and a fault prognosis indicator (FPI) to trigger alarms in advance due to the main bearing fault evolution in time.

A. GRU Proposed Architecture

A GRU neural network is employed in this study to build the normality model. Despite the intention of this subsection is not extensively delve into GRUs, a brief explanation is given and an overall justification on why these are used instead of LSTMs, RNNs, or ANNs is given also.

GRUs have two gates: update and reset. In short, these two vectors decide what information should be passed or not to the output. The key point is that they are trained to retain long-term information without washing it through time and remove information which is irrelevant for the forecasting, see [11], [12], [13], [14].

RNNs are the most basic architectures that pay attention to the past for predicting the future. However, they have the serious vanishing gradient problem. Based on that, they may not be appropriate to work on this study. On the other hand, compared to LSTMs [15], GRUs only have two gating signals, whereas LSTMs have three. Thus, GRUs have fewer parameters and lower computational cost.

In prognosis models, learning how past data samples affect the future data samples is crucial. Nevertheless, understanding the nature of ANNs, they do not consider previous data to predict future data, thus, neither these neural networks are not suitable to address the goal of the study. In a nutshell,

Table II. Setup of the GRU hyperparameters.

Hyperparameter	Value
Number of hidden layers	1
Number of neurons in the hidden state	128
Epoch size	50
Batch size	128
Initial learning rate	0.001
Loss function	MSE

with this brief discussion, GRUs demonstrate to be the most appropriate to deal with the needs of the work.

The GRU proposed is based on a many-to-one structure, where the output is the main bearing temperature while the inputs are the other variables are: generator bearing temperature, gearbox oil temperature, and the primary wind speed. The inputs are given in a sequence of 144 samples, considering a window length of 24 hours (1 sample/10-min x 144 10-min/day = 144 samples/day).

All deep learning models require a set of hyperparameters for tuning and improving the performance according to the data. A GRU is not the exception, and in Table II are shown the hyperparameters configured. Each one is very important and is selected after proving several values until reaching a good performance.

B. How is the FPI structured?

A typical FPI is based on a threshold, i.e. depending on the setup of the threshold, if the residual exceeds it or not an alarm is triggered. In this study specifically, the residual is defined as the absolute difference between the real main bearing temperature (T) and the predicted by the GRU model ($\hat{T}$). Nevertheless, if this raw residual is used, a large number of false positives would be generated, thus it is necessary to early avoid this problem.

If there is a truly abnormal behavior a notable number of continuous raw residuals have to exceed the threshold, thus, by applying a strategy to condense or average the raw residuals through time, the persistence of the abnormal behavior will remain notable. In this work particularly, a technique based on moving average (MA) [16] is used to smooth the trend of the raw residuals.

The basic MA may produce cyclic and trend-like plots, regardless if the original data are themselves independent shuffle events with a stationary mean [17]. To deal with that, the exponentially weighted MA (EWMA) appears. This technique assigns less weights as the data get older. In short, the EWMA is expressed as follows:

$$\text{EWMA} = \hat{T}_t + \lambda(T_t - \hat{T}_t), \tag{1}$$

where $\hat{T}_t$ is the predicted value at time t, T_t is the measured real SCADA value at time t, and λ is a parameter ($0 < \lambda < 1$) that defines the memory depth of the EWMA. This parameter is empirically selected using its relation to the span ($s \geq 1$) [18]. Equation 2 shows how these parameters are related:

$$\lambda = \frac{2}{s+1}. \tag{2}$$

After dealing with the possible problems of the raw residuals, it is essential to introduce the concept of the used threshold. This is a limit to monitor if the residual is within normal (healthy) behavior or not. For that, the residuals of tested training-validation-data are used. Likewise, the mean (μ) and the standard deviation (σ) are employed to build the threshold's equation (Equation 3) applied to those residuals.

$$\text{threshold} = \mu + \kappa\sigma, \tag{3}$$

where κ is a weight that states the threshold value. In this study, two κ-values are used, as one defines a warning alarm and the other one a fault alarm as such.

Finally, Figure 2 summarizes in a flowchart all steps discussed in the data-preprocessing section and those related to the fault prognosis methodology.

5. Results and Discussion

This section begins showing the residuals based on the absolute difference (Figure 3) for the two WTs considered in this study, where WT1 is healthy and WT2 is faulty. Recall that the tag "healthy" supposes that in WT1's test-dataset there is no any presence of the main bearing fault while the tag "faulty" suggests in WT2's test-dataset there is presence of the main bearing fault, in this case, occurred on June 11, 2018. Figure 3 (a) and (b) reports the residuals for the train and test datasets of WT1 while Figure 3 (c) and (d) reports the same but for WT2.

Looking at WT1's residuals, it is clear that they do not exceed 0.12 against the WT2's residuals, where the limit is set in 0.20. Furthermore, it is clearly visible that before the main bearing fault there is an abnormal trend of the residuals suggesting something is happening some months in advance. After the main bearing is changed, the residuals are lower, which indicates all revert back to normal.

As can be seen in Figure 3 the residuals show a lot of peaks which many times generate false alarms. In the last section, it was introduced the use of the EWMA as a technique to reduce false alarms and extract the trend of the data as a function of persistence through time. The most important factor - to carry the EWMA out - defined as s and called "spans" indicates the memory depth of the EWMA. Several values of span were proven to measure the performance of the EWMA on the residuals' processing. Among the considered spans are: $s = 144$ (one day), $s = 1008$ (one week), and $s = 2016$ (two weeks). In this work, the results on $s = 2016$ are reported, since they demonstrated being more robust to false alarms. Additionally, even when the EMWA's memory-depth is based on two weeks, the abnormal trend because of the main bearing fault is very visible and persistent through time, i.e., several months in advance.

After computing the EMWA, the definition of the threshold is stated. As it was explained in last section, the purpose of these thresholds is working as a warning and a fault alarm as such. Recall that in Equation 3 the value of κ establishes how the threshold varies, thus, in this case, two values are proposed: 12 and 15. Likewise, the results showed that with

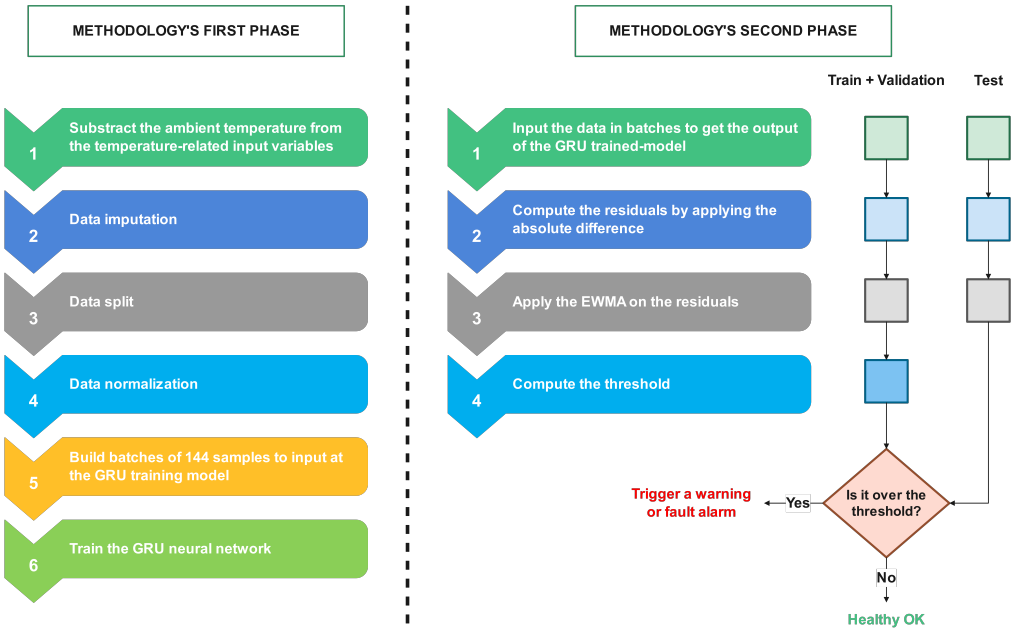

Fig. 2: Flowchart of the proposed methodology where the first phase involves all the related to data preprocessing and training of GRU. Next, the second phase takes place after getting the output from the GRU trained-model.

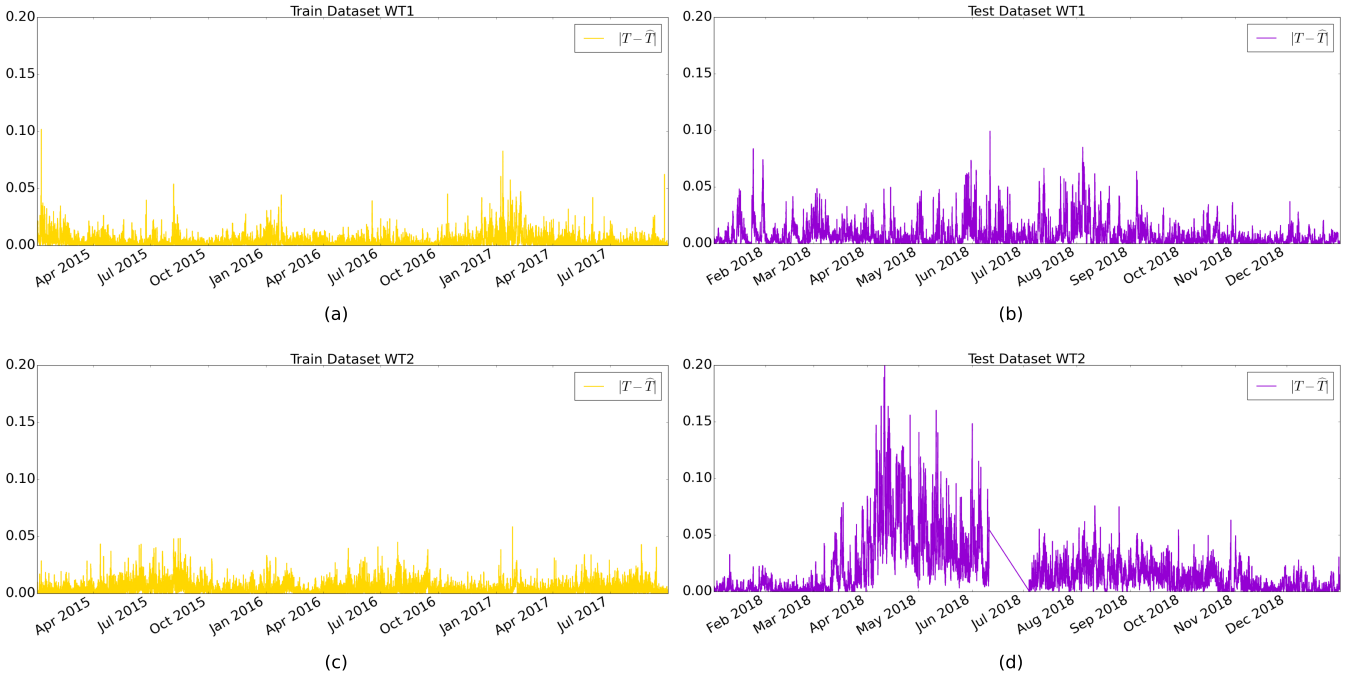

Fig. 3: Residuals based on the absolute difference, $|T - \hat{T}|$, for both WT1 and WT2. (a) and (b) are computed on the train dataset while (c) and (d) are computed on the test dataset.

these values together with the EWMA computation build a robust framework to avoid the false alarms and be effective in the detection and prognosis of the main bearing fault on any WT.

Figure 4 finally shows the results for WT1 and WT2. As it was alerted, there are two thresholds: one set in $\mu + 12\sigma$ (green dotted line) that defines the warnings and the other one set in $\mu + 15\sigma$ (red dotted line) that indicates a fault alarm. For WT1, at any moment the residuals exceed the thresholds which reports there is no any indication of the main bearing fault. On the other hand, for WT2 it is clear that residuals exceed both thresholds in the first week in April, 2018. Recalling that the main bearing fault occurred on June 11, 2018, the prognosis is done two months in advance at least.

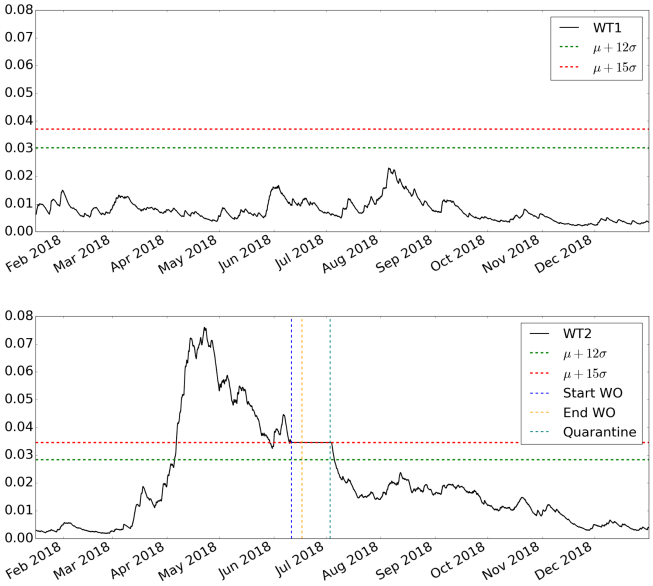

Fig. 4: FPI methodology applied on the processed residuals with the EWMA. The green dotted-line indicates a warning, while the red dotted-line indicates a fault alarm as such.

6. Conclusions

By using a GRU neural network and just employing real SCADA data an early fault detection methodology is proposed and developed in this work. The model is constructed from scratch, only with healthy data, i.e., establishing a normality (normal behavior) model. The results are reported on two WTs, one being healthy and the other one being faulty. Through the fault prognosis methodology, the results demonstrate that the model is robust and does not generate false alarms while the fault prognosis is effectively done two months in advance at least. In other words, the fault prognosis methodology shows that on the healthy WT none of the two thresholds (warning and fault) are exceeded. On the other hand, on the faulty WT both threshold are exceeded which clearly indicates there is an abnormal behavior.

Thus, this early prognosis clearly offers a strategy for WTs' operators to plan ahead for repairs and maintenance orders, decreasing the downtime of WTs when these finally fault and corrective maintenance is mandatory. Instead of that, the intention is increasing the uptime of the WTs as much as possible and extend the lifespan of the WTs' components.

References

[1] I. E. A. , Offshore wind outlook 2019. world energy outlook special report. (2019).

[2] T. Stehly, P. Beiter, P. Duffy, 2019 cost of wind energy review-doi:https://dx.doi.org/10.2172/1756710.

[3] J. Carroll, A. McDonald, D. McMillan, Failure rate, repair time and unscheduled o&m cost analysis of offshore wind turbines, Wind Energy 19 (6) (2016) 1107–1119. doi:https://doi.org/10.1002/we.1887.

[4] J. J. Nielsen, J. D. Sørensen, On risk-based operation and maintenance of offshore wind turbine components, Reliability Engineering System Safety 96 (1) (2011) 218–229, special Issue on Safecomp 2008. doi:https://doi.org/10.1016/j.ress.2010.07.007.

[5] H. ZHAO, Y. GAO, H. LIU, L. Li, Fault diagnosis of wind turbine bearing based on stochastic subspace identification and multi-kernel support vector machine, Journal of Modern Power Systems and Clean Energy 7 (2019) 350–356. doi:10.1007/s40565-018-0402-8.

[6] Y. Fu, Z. Gao, A. Zhang, X. Liu, Fault classification for wind turbine benchmark model based on hilbert-huang transformation and support vector machine strategies, in: 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), 2021, pp. 1–8. doi:10.1109/INDIN45523.2021.9557362.

[7] J. Tautz-Weinert, S. Watson, Using scada data for wind turbine condition monitoring - a review, IET Renewable Power Generation 11 (2017) 382–394. doi:10.1049/iet-rpg.2016.0248.

[8] Encalada-Dávila, B. Puruncajas, C. Tutivén, Y. Vidal, Wind turbine main bearing fault prognosis based solely on scada data, Sensors 21 (6) (2021). doi:10.3390/s21062228.
URL https://www.mdpi.com/1424-8220/21/6/2228

[9] Z. Jiang, W. Hu, W. Dong, Z. Gao, Z. Ren, Structural reliability analysis of wind turbines: A review, Energies 10 (12) (2017). doi:10.3390/en10122099.
URL https://www.mdpi.com/1996-1073/10/12/2099

[10] Z. Zhang, Automatic fault prediction of wind turbine main bearing based on scada data and artificial neural network, Open Journal of Applied Sciences 8 (2018) 211–225. doi:10.4236/ojapps.2018.86018.

[11] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O'Reilly Media, Sebastopol, CA, 2019.

[12] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, CoRR abs/1412.3555 (2014). arXiv:1412.3555.
URL http://arxiv.org/abs/1412.3555

[13] Z. Wu, S. King, Investigating gated recurrent networks for speech synthesis, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5140–5144. doi:10.1109/ICASSP.2016.7472657.

[14] R. Dey, F. M. Salem, Gate-variants of gated recurrent unit (gru) neural networks, in: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017, pp. 1597–1600. doi:10.1109/MWSCAS.2017.8053243.

[15] A. Flores, H. Tito-Chura, V. Yana-Mamani, Wind speed time series prediction with deep learning and data augmentation, in: K. Arai (Ed.), Intelligent Systems and Applications, Springer International Publishing, Cham, 2022, pp. 330–343.

[16] J. S. Hunter, The exponentially weighted moving average, Journal of Quality Technology 18 (4) (1986) 203–210. doi:10.1080/00224065.1986.11979014.

[17] L. S. Nelson, The deceptiveness of moving averages, Journal of Quality Technology 15 (2) (1983) 99–100. doi:10.1080/00224065.1983.11978852.

[18] P. Cisar, S. M. Cisar, Ewma statistic in adaptive threshold algorithm, in: 2007 11th International Conference on Intelligent Engineering Systems, 2007, pp. 51–54. doi:10.1109/INES.2007.4283671.