

Wind Turbine Gearbox Condition Monitoring using Normal Behavior Modeling based on lubrication Pressure

Yazid Aafif

Nancy Research Center in Automatic Control (CRAN), UMR 7039
University of Lorraine
Nancy, France
Yazid.aafif@univ-lorraine.fr

Anis Chelbi

Department of Mechanical Engineering
RIFTSI Lab: Intelligent Robotics, Reliability, Signal and Image Processing
ENSIT, University of Tunis, Tunis, Tunisia
Mediterranean School of Business, South Mediterranean University
Tunis, Tunisia
Anis.chelbi@planet.tn

Sofiène Dellagi

LGIPM, Computer Engineering, Production and Maintenance Laboratory
University of Lorraine
Metz, France
Sofiène.dellagi@univ-lorraine.fr

Abstract

This paper describes a machine learning framework for gearbox health assessment in wind turbines, built around gearbox oil pump pressure rather than the thermal signals that dominate the existing literature. A Random Forest Normal Behavior Model (NBM) is used based on a healthy operational baseline to predict expected lubrication pressure, and the residual between predicted and actual values is smoothed over a 30-day rolling window to filter out operational variability and expose structural drift. The model was chosen after benchmarking against XGBoost and Support Vector Regression.

Keywords

Wind turbine, gearbox condition monitoring, SCADA, normal behavior modeling, Random Forest.

1. Introduction

Wind energy now accounts for a significant share of electricity production in many countries. As turbine fleets age, operation and maintenance (O&M) costs keep growing. Among turbine components, the gearbox is one of the costliest to maintain (Aafif et al. (2024)).

Industry data put the gearbox failure rate at about 0.154 events per turbine per year, with an average downtime of 14 to 15 days and repair costs that can exceed \$250,000 per event (Sheng 2013, Dao et al. 2019). Most of these failures

originate from the bearings, and they are typically preceded by a gradual temperature rise. Aafif et al. (2022) showed that in practice, gearbox condition is monitored through temperature: when a threshold is exceeded, production is reduced to let the gearbox cool before resuming normal operation. Detecting this rise early enough is what makes thermal monitoring useful.

SCADA systems record operational data at 10-minute intervals on virtually all modern turbines. The data is already there, collected continuously at no additional cost. Several studies have used it to develop data-driven fault detection models, with results generally showing that deviations from normal thermal behavior can be caught well before a failure occurs (Jankauskas et al. 2023, Leahy et al. 2019, Jiang et al. 2022).

Most SCADA-based condition monitoring works use thermal signals; lubrication pressure has received little attention. A Random Forest regression model is trained on a baseline corresponding to a healthy state to capture normal lubrication system behavior, and the residual between observed and predicted pressure is smoothed over a 30-day rolling window.

The main contributions of this work are: (i) a multi-year SCADA harmonization procedure addressing inter-annual parameter inconsistencies across six annual datasets; (ii) a physics-informed zero-pressure artifact removal step specific to the lubrication circuit and (iii) a Random Forest-based normal behavior model for gear oil pump pressure.

2. Literature Review

Condition monitoring of wind turbine gearboxes using SCADA data has been extensively studied over the past decade. Tautz-Weinert and Watson (2017), Pandit et al. (2023), and Pandit and Wang (2024) all point to Normal Behavior Modeling (NBM) as the prevailing approach: train a model on healthy operation data, then monitor the residual between predicted and actual measurements over time. When that residual drifts persistently, the component is no longer behaving as the model expects.

Within this NBM framework, the choice of monitored signal has been almost uniformly restricted to thermal variables. Early contributions such as Wilkinson et al. (2014) used polynomial regression on gearbox and generator temperatures and demonstrated fault detection between one month and two years in advance. Bangalore and Tjernberg (2020) applied Support Vector Regression to gearbox bearing temperatures, using Mahalanobis distance for anomaly scoring. Yuan et al. (2019) took a stacking ensemble approach, combining Random Forest, Gradient Boosted Decision Trees, and XGBoost to model gearbox oil temperature, with Mahalanobis distance replacing the direct residual as the deviation metric. More recent work has moved toward deep learning: Oliveira-Filho et al. (2025) applied autoencoders, LSTMs, and variational autoencoders to SCADA temperature time series, while Puruncajas et al. (2025) used artificial neural networks to detect bearing temperature anomalies across a multi-turbine wind farm. Across all these works, the monitored signal is a temperature, and the output is either a binary alarm or a residual trend without normalization into a bounded health indicator.

The treatment of gearbox pressure signals in this context is notably absent. To the best of our knowledge, the only directly adjacent work is Zeng et al. (2020), who applied Sparse Bayesian Learning to model the differential pressure across the gearbox oil filter and used hypothesis testing to flag anomalies. However, that work monitors pressure difference across a filter rather than absolute pump pressure. It targets a distinct physical failure mode related to filter clogging, and produces a probabilistic alarm flag rather than a continuous health index. Milani et al. (2025) included pressure signals as one of several SCADA variables in a generative model for synthetic data augmentation, but pressure there plays no role as a health indicator.

Gear oil pump pressure reflects the mechanical state of the lubrication system directly. Yet no prior work has applied a Normal Behavior Modeling framework to this signal to extract a load-decoupled degradation indicator. The present work addresses this gap: a Random Forest model is trained on a healthy operational baseline to predict expected lubrication pressure from eight operational and environmental inputs, and the residual between predicted and actual values is smoothed over a 30-day rolling window to isolate structural drift from operational variability, using only standard SCADA instrumentation.

3. Data Aggregation and Harmonization

3.1 Data Description

The study is based on multi-year SCADA data collected from a single wind turbine. The dataset comes from the Penmanshiel Wind Farm SCADA system in the United Kingdom, a publicly available resource (Plumley (2022)). Measurements were recorded every 10 minutes between June 2016 and June 2021, giving roughly 263,340 records across 300 operational channels. Six annual datasets were initially available, each corresponding to a different observation period. Because the recorded parameters slightly differed from one year to another, a harmonization step was first conducted.

To keep only the variables that were shared across all datasets, a column intersection procedure was used. By doing this, temporal consistency was guaranteed, and structural biases that might result from parameters that only appeared in particular years were avoided. The resulting unified dataset contains a stable set of operational, thermal, electrical, and drivetrain-related variables spanning the full analysis horizon.

3.2 Missing Values

To guarantee data quality, columns containing more than 70% missing values were removed before any modeling step, in line with standard preprocessing practice (Chesterman et al., 2023; Pandit et al., 2023). Missing entries were defined as NaN (Not a Number) values, empty cells, or zero placeholders when physically implausible.

A year-by-year breakdown of gear oil pump pressure (Table 1) revealed a sharp discontinuity in the raw signal. In 2016 and 2017, mean pressure sits around 5 bar and never exceeds 13 bar. From 2018 onward it rises to 342 bar, stabilizing around 540 to 560 bar through 2021. A difference of two orders of magnitude rules out degradation or seasonality as explanations. The most likely cause is a sensor configuration change or a different physical quantity being logged in the early years, consistent with the inter-annual parameter mismatches found during the harmonization step and with data quality issues documented for publicly available SCADA datasets (Leahy et al., 2019). Records from 2016 and 2017 were therefore excluded entirely.

Table 1: Gear oil pump pressure descriptive statistics by year (raw dataset)

Year	Count	Mean (bar)	Std (bar)	Min (bar)	Median (bar)	Max (bar)
2016	28,943	5.1	2.0	0.0	5.3	12.7
2017	52,394	5.7	1.6	0.0	5.7	9.4
2018	52,456	342.7	289.6	0.1	466.8	888.1
2019	52,174	549.3	175.6	0.1	549.1	1140.2
2020	49,604	562.8	175.5	2.8	558.6	1029.2
2021	26,063	541.8	187.0	0.6	541.6	955.4

Within the 2018 to 2021 period, residual near-zero readings were also removed. Sensors during turbine standstill frequently record default rather than operational values, and filtering these periods out is standard practice in SCADA-based condition monitoring (Letzgus, 2020; Tautz-Weinert and Watson, 2017). A 50 bar threshold was applied on physical grounds: a running turbine cannot sustain lubrication pressure below that level. The valid operational dataset after filtering comprises 156,567 records spanning April 2018 to June 2021.

4. Construction of the Normal Behavior Model (NBM)

4.1 Normal Behavior Modeling

Raw condition indicators obtained from SCADA systems, such as gear oil pump pressure, are intrinsically coupled with operational and environmental variables. In wind turbines, lubrication pressure does not evolve independently; it varies with wind speed, rotor speed, electrical loading, and thermal conditions. Consequently, direct monitoring of raw pressure values does not allow a clear distinction between variations induced by operating conditions and those resulting from mechanical degradation.

For example, an increase in wind speed leads to increased mechanical loading, which may cause a rise in pressure. Without correcting for this load dependence, such variation could be incorrectly interpreted as degradation. Therefore, a separation between operational variability and structural deterioration is required.

To address this issue, a normal behavior modeling framework was adopted, consistent with established practice in SCADA-based wind turbine condition monitoring (Chesterman et al., 2023; Wilkinson et al., 2014).

Let:

$$\begin{aligned} Y(t) &= \text{Gear oil pump pressure at time } t \\ X(t) &= \text{Vector of operational and environmental variables at time } t \end{aligned}$$

The pressure signal can be conceptually decomposed as:

$$Y(t) = f(X(t)) + \delta(t) \quad \text{Eq (1)}$$

where:

$f(X(t))$ represents the expected healthy behavior of the gearbox under given operating conditions,
 $\delta(t)$ captures abnormal deviations attributable to degradation or mechanical anomalies.

The objective is to estimate $f(\cdot)$ so that the degradation component $\delta(t)$ can be isolated.

4.2 Baseline Selection and Model Training

The training baseline was the first 20% of the cleaned records in chronological order, covering April 2018 to December 2018 (31,313 records at 10-minute intervals). In practice, identifying a verified fault-free baseline from operational SCADA data is rarely possible: real machines carry no label indicating whether a given record reflects healthy operation or not (Chesterman et al., 2023; Letzgus, 2020). The operational logbook (Plumley, 2022) shows the turbine ran at full performance for 60.1% of the baseline period. The remainder is accounted for by environmental standstill, technical standby, and routine maintenance visits. Nothing in the logbook points to lubrication system degradation during this window. Five hydraulic oil flushing operations were recorded as technical standby reflecting scheduled circuit maintenance. The one gearbox speed warning in November 2018 is a transient sensor event with no connection to lubrication pressure. Taken together, the logbook provides no grounds to suspect the baseline is contaminated by progressive lubrication circuit deterioration.

The eight input variables were wind speed, electrical power, rotor speed, generator RPM, ambient temperature, nacelle temperature, drivetrain acceleration, and grid current, selected for physical relevance and confirmed through correlation analysis against gearbox pressure.

Three candidate models were evaluated on the baseline dataset: Random Forest, XGBoost, and Support Vector Regression.

Random Forest achieved the highest R^2 of 0.79 with a MAE of 52 bar, marginally outperforming XGBoost ($R^2 = 0.79$, MAE = 52.73 bar) and substantially outperforming SVR ($R^2 = -0.34$), which failed to generalize due to the high dimensionality and variance of the SCADA pressure signal. Random Forest was therefore selected as the normal behavior model. Although an R^2 of 0.79 indicates that approximately 21% of the pressure variance remains unexplained, this residual uncertainty is attributable to sensor noise and short-term mechanical dynamics that no 10-minute averaged SCADA variable can fully capture. For the purpose of health index construction, this level of fitness is sufficient: the model successfully removes the dominant operational component from the pressure signal. The remaining unexplained variance is subsequently attenuated by the 30-day rolling mean smoothing applied to the residual in the following section.

Random Forest was configured with 100 estimators, a maximum tree depth of 20, and a minimum of 30 samples per leaf.

The trained model provides an estimate of expected healthy pressure:

$$\hat{Y}(t) = f(X(t)) \quad \text{Eq (2)}$$

where $\hat{Y}(t)$ denotes the predicted pressure under healthy conditions.

Figure 1 compares the raw gear oil pump pressure with the Random Forest model's predictions. The predicted signal tracks the measured values closely while smoothing out spikes and sensor artifacts. The fit confirms that the eight input variables carry enough information to reconstruct expected pressure $\hat{Y}(t)$ under normal operating conditions.

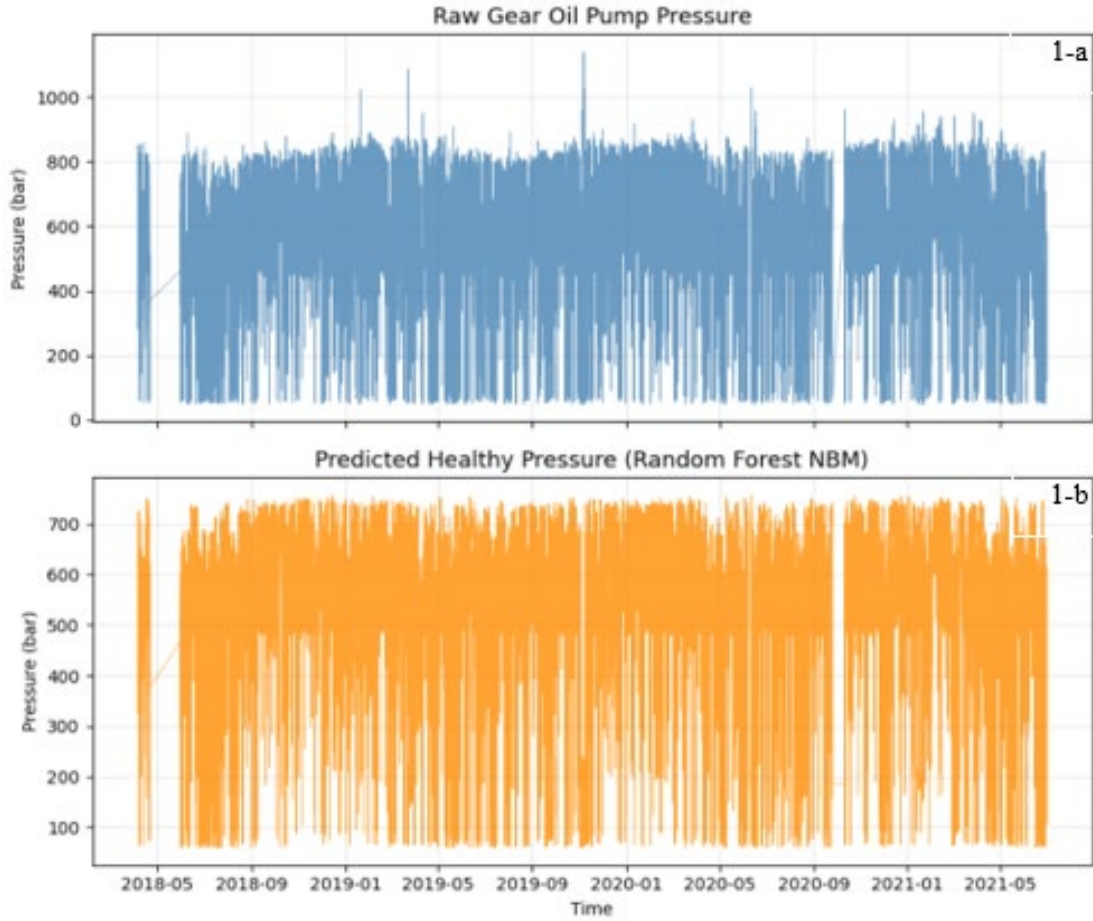


Figure 1: Raw gear oil pump pressure (Figure 1-a) and Random Forest predicted healthy pressure (Figure 1-b) over the full observation period

4.3 Residual Construction and Degradation Signal

Once the model was trained, predictions were generated for the entire dataset (including post-baseline periods) (Figure 1-b). For each timestamp, the degradation signal was defined as the residual:

$$r(t) = Y(t) - \hat{Y}(t) \quad \text{Eq (3)}$$

This residual quantifies the deviation of actual pressure from expected healthy behavior after accounting for operational influence.

If the gearbox remains healthy, the residual should fluctuate around zero with no trend. Conversely, a persistent drift in residual magnitude indicates progressive degradation.

Figure 2 shows the raw residual signal after subtracting the model prediction from the actual pressure. The residual fluctuates around zero, confirming that the normal behavior model has successfully removed the operational component. However, the high-frequency noise makes it impossible to identify any degradation trend directly.

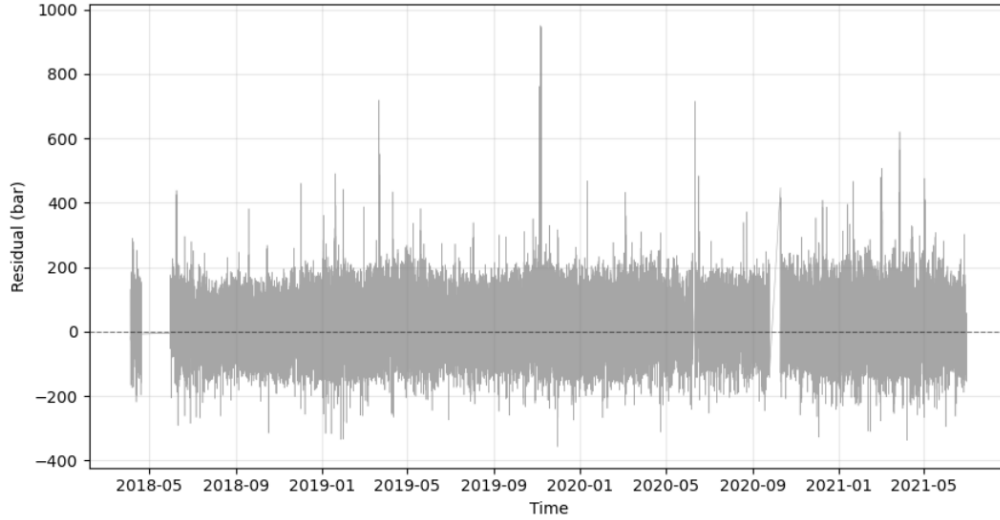


Figure 2: Residual Signal Representing Gearbox Degradation Indicator

4.4 Long-Term Residual Smoothing

SCADA data exhibit high-frequency variability due to rapid fluctuations in load and environmental conditions. To isolate structural degradation from short-term operational noise, a rolling average was applied to the residual signal ((Qiu et al., 2012; Wilkinson et al., 2014)).

A 30-day rolling average was computed:

$$r_{smooth}(t) = \frac{1}{W} \sum_{i=t-W+1}^t r(i) \quad \text{Eq (4)}$$

where W corresponds to a 30-day window (approximately 4320 samples for 10-minute spaced SCADA data).

This smoothing window was selected based on the assumption that mechanical degradation evolves at a slower time scale than operational fluctuations. The objective was to extract progressive structural drift rather than short-term anomalies.

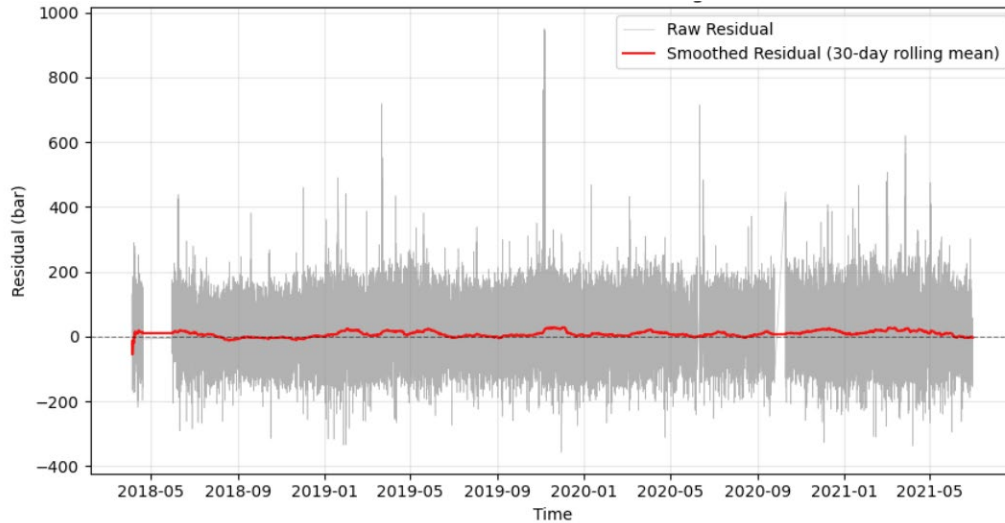


Figure 3: Smoothed Residual Signal Representing Gearbox Degradation Indicator

Figure 3 shows the smoothed residual over the full observation period. While the rolling mean successfully suppresses short-term noise and exposes structural drift, the resulting signal remains expressed in physical units of pressure and its magnitude depends on the residual variance of the specific turbine and operating regime.

6. Conclusion

This paper described a machine learning framework for gearbox health assessment in a wind turbine, using gear oil pump pressure drawn from a raw dataset spanning six years of SCADA data, of which three years of valid operational records were retained after physics-informed artifact removal. A Random Forest model was trained on a healthy baseline to relate eight operational and environmental inputs to expected lubrication pressure under normal conditions. It outperformed XGBoost and Support Vector Regression, reaching an R^2 of 0.79. The residual between predicted and actual pressure was smoothed with a 30-day rolling mean, yielding a clean degradation signal.

The construction of a normalized, bounded Gearbox Health Index from this smoothed residual is the subject of ongoing work.

References

- Aafif, Y., Chelbi, A., Mifdal, L., Dellagi, S. and Majdouline, I., Optimal preventive maintenance strategies for a wind turbine gearbox, *Energy Reports*, vol. 8, pp. 803–814, 2022.
- Aafif, Y., Schutz, J., Dellagi, S., Chelbi, A. and Mifdal, L., Integrated design and maintenance strategies for wind turbine gearboxes, *Journal of Quality in Maintenance Engineering*, vol. 30, no. 3, pp. 521–539, 2024.
- Bangalore, P. and Tjernberg, L. B., Wind turbine gearbox condition monitoring based on class of support vector regression models and residual analysis, *Sensors*, vol. 20, no. 23, pp. 6742, 2020.
- Chesterman, X., Verstraeten, T., Daems, P.-J., Nowé, A. and Helsen, J., Overview of normal behavior modeling approaches for SCADA-based wind turbine condition monitoring demonstrated on data from operational wind farms, *Wind Energy Science*, vol. 8, pp. 893–924, 2023.
- Dao, C., Kazemtabrizi, B. and Crabtree, C., Wind turbine reliability data review and impacts on levelised cost of energy, *Wind Energy*, vol. 22, no. 12, pp. 1848–1871, 2019.
- Jankauskas, M., Serackis, A. and Sapurov, M., Exploring the limits of early predictive maintenance in wind turbines applying an anomaly detection technique, *Sensors*, vol. 23, no. 12, pp. 5695, 2023.
- Jiang, G., Li, W., Fan, W., He, Q. and Xie, P., TempGNN: a temperature-based graph neural network model for system-level monitoring of wind turbines with SCADA data, *IEEE Sensors Journal*, vol. 22, no. 23, pp. 22894–22907, 2022.
- Leahy, K., Gallagher, C., O'Donovan, P. and O'Sullivan, D. T. J., Issues with data quality for wind turbine condition monitoring and reliability analyses, *Energies*, vol. 12, no. 2, pp. 201, 2019.
- Letzgus, S., Change-point detection in wind turbine SCADA data for robust condition monitoring with normal behaviour models, *Wind Energy Science*, vol. 5, no. 4, pp. 1375–1397, 2020.
- Milani, A. E., Zappala, D., Castellani, F. and Watson, S., Simulating run-to-failure SCADA time series to enhance wind turbine fault detection and prognosis, *Wind Energy Science*, vol. 10, pp. 2563–2576, 2025.
- Oliveira-Filho, A., Zemouri, R., Cambron, P. and Tahan, A., Wind turbine SCADA data imbalance: a review of its impact on health condition analyses and mitigation strategies, *Energies*, vol. 18, no. 1, pp. 59, 2025.
- Pandit, R. and Wang, J., A comprehensive review on enhancing wind turbine applications with advanced SCADA data analytics and practical insights, *IET Renewable Power Generation*, vol. 18, no. 4, pp. 722–742, 2024.
- Pandit, R. K., Astolfi, D., Hong, J., Infield, D. and Santos, M., SCADA data for wind turbine data-driven condition/performance monitoring: a review on state-of-art, challenges and future trends, *Wind Engineering*, vol. 47, no. 2, pp. 422–441, 2023.
- Plumley, C., Penmanshiel wind farm data (0.0.2) [Data set], Zenodo, <https://doi.org/10.5281/zenodo.5946808>, 2022.
- Puruncajas, B., Encalada-Davila, A., Tutiven, C. and Vidal, Y., Use of artificial neural networks and SCADA data for early detection of wind turbine gearbox failures, *Machines*, vol. 13, no. 8, pp. 746, 2025.
- Qiu, Y., Feng, Y., Tavner, P., Richardson, P., Erdos, G. and Chen, B., Wind turbine SCADA alarm analysis for improving reliability, *Wind Energy*, vol. 15, no. 8, pp. 951–966, 2012.
- Rahman, M. A., Sarker, B. R. and Escobar, L. A., Peak demand forecasting for a seasonal product using Bayesian approach, *Journal of the Operational Research Society*, vol. 62, pp. 1019–1028, 2011.
- Sheng, S., Report on wind turbine subsystem reliability: a survey of various databases, National Renewable Energy Laboratory, NREL/PR-5000-59111, 2013.

- Tautz-Weinert, J. and Watson, S. J., Using SCADA data for wind turbine condition monitoring: a review, IET Renewable Power Generation, vol. 11, no. 4, pp. 382–394, 2017.
- Wilkinson, M., Darnell, B., Van Delft, T. and Harman, K., Comparison of methods for wind turbine condition monitoring with SCADA data, IET Renewable Power Generation, vol. 8, no. 4, pp. 390–397, 2014.
- Yuan, T., Sun, Z. and Ma, S., Gearbox fault prediction of wind turbines based on a stacking model and change-point detection, Energies, vol. 12, no. 22, pp. 4224, 2019.
- Zeng, X., Yang, M. and Bo, Y., Gearbox oil temperature anomaly detection for wind turbine based on sparse Bayesian probability estimation, International Journal of Electrical Power and Energy Systems, vol. 123, pp. 106233, 2020.

Biographies

Yazid AAFIF is a doctor-engineer and currently serves as an ATER (Temporary Teaching and Research Associate) at Polytech Nancy, University of Lorraine, France. He completed his Ph.D. in Automatic Control, Signal Processing, and Computer Engineering at the LGIPM laboratory (University of Lorraine, Metz), specializing in the optimization of maintenance strategies for wind turbines. He also holds a Master's degree in Complex Systems Engineering (Industrial Engineering & Logistics) from the University of Lorraine, and a State Engineer's degree in Industrial Engineering from the École Polytechnique d'Agadir, Morocco. His research focuses on developing optimal preventive and predictive maintenance strategies for critical industrial systems, particularly wind turbines, integrating reliability modeling, economic constraints, and operational data analytics including SCADA-based machine learning approaches. He has authored several publications in international journals such as Computers & Industrial Engineering and Energy Reports, and has presented his work at numerous international conferences.

Anis CHELBI is a full Professor in Industrial Engineering at the Higher National School of Engineering at the University of Tunis. He is also Visiting Professor at the Mediterranean School of Business in Tunis. His main expertise and research interests include sustainable supply chains management and systems maintenance and reliability. Prof. Chelbi heads the RIFTSI (Intelligent Robotics, Reliability and Signal and Image Processing) research laboratory. He is also an associate member of the Interuniversity Research Center on Enterprise Networks, Logistics and Transportation (CIRRELT) in Canada. In addition to his numerous scientific publications, he has over 25 years' experience as a consultant and trainer for national and international companies in various fields, including the aerospace industry with SNECMA Motors and Messier-Dowty in France, the chemical industry with AKZO-Nobel in Belgium, and other industries in Tunisia such as ceramics, packaging and agri-food. Professor Chelbi. He has also been Associate Editor of the scientific journal Heliyon Engineering from 2021 to 2023, and is currently Associate Editor of the Journal of Quality in Maintenance Engineering. Finally, he is the recipient of the Outstanding Paper Award at the Literati Network Awards for Excellence, 2009.

Sofiène DELLAGI is a full Professor in Industrial Engineering. He received his Ph.D. degree in Industrial Engineering from the University of Lorraine, France. He has been an Associate Professor in Industrial Engineering with the University of Lorraine since 2008, and director of LGIPM Laboratory, since 2024. His research has led to active partnerships with industrial players, including three R&D contracts (SFER and ArcelorMittal Research Center), and strong international academic collaborations with institutions such as the École de Technologie Supérieure de Montréal, the École Supérieure de Sciences et Techniques de Tunis, and Université Laval, Québec. He has co-edited a special issue of the International Journal of Production Research (2008) and has served as a reviewer and jury member for numerous international conferences and doctoral defenses. Prof. Dellagi has received several distinctions, including the Lorraine Regional Councils Prize for his doctoral thesis, the doctoral supervision bonus (PEDR, 2018–2022), and a full-year research leave (CRCT) awarded by the CNU in 2016–2017.