UNIVERSITAT POLITÉCNICA DE CATALUNYA

UNITAT TRANSVERSAL DE GESTIÓ DE L'ÀMBIT DE CAMINS

DOCTORAL THESIS

ENVIRONMENTAL ENGINEERING

Use of Advanced Analytics for Health Estimation and Failure Prediction of Wind Turbines

Author: Mattia Beretta

Director: Jordi Cusidò Roura

Barcelona, April 2022

Thesis by compendium of publications

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Use of Advanced Analytics for Health Estimation and Failure Prediction of Wind Turbines

by
Mattia Beretta

Doctoral Thesis submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy in Environmental Engineering

Director: Jordi Cusidò Roura

Barcelona, Sunday 3rd April, 2022

Thesis by compendium of publications

Unitat Transversal de Gestió de l'Àmbit de Camins
c/ Jordi Girona, 1-3
08034, Barcelona, España

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Resum

El sector energètic ha experimentat importants canvis i revolucions en les últimes dècades. Les fonts d'energia renovables han crescut significativament, i ara representen una part important en el conjunt de generació. L'energia eòlica ha augmentat significativament, convertint-se en una de les millors alternatives per produir energia verda. La recerca i la innovació ha ajudat a reduir considerablement els costos de producció i operació de l'energia eòlica, però encara hi ha oberts reptes importants. Aquesta tesi aborda el manteniment predictiu i el seguiment d'aerogeneradors, amb l'objectiu de presentar solucions d'algoritmes de predicció dissenyats tenint en compte les necessitats de la indústria. Més concretament conceptes com, la **interpretabilitat**, **escalabilitat**, **modularitat** i **fiabilitat** de les prediccions ho són els objectius, juntament amb els requisits **limitats per les de dades** disponibles d'aquest projecte. De totes les dades disponibles a disposició dels operadors d'aerogeneradors, les dades del sistema SCADA són la principal font d'informació utilitzada en aquest projecte, per la seva àmplia disponibilitat i baix cost. En el present treball, els models de conjunt tenen un paper important en el desenvolupament dels marcs predictius presentats gràcies al seu caràcter modular que permet l'ús d'algoritmes i tipus de dades molt diversos. Resultats importants obtinguts d'aquests experiments són l'efecte beneficiós de combinar múltiples i diverses fonts de dades, per exemple, SCADA i dades d'alarmes, la facilitat de combinar diferents algorismes i indicadors i el notable guany en predir el rendiment que es pot oferir. Finalment, donat el paper central que SCADA l'anàlisi de dades juga en aquesta tesi, però també en la indústria de l'energia eòlica, una anàlisi detallada de la es presenten les limitacions i les mancances de les dades SCADA. En particular es va estudiar l'efecte de l'agregació de dades —una pràctica habitual en la indústria eòlica—. Dins d'aquest treball es proposa un marc metodològic que s'ha utilitzat per estudiar dades SCADA d'alta freqüència. Això va portar a la conclusió que els períodes d'agregació típics, de 5 a 10 minuts que són l'estàndard a la indústria de l'energia eòlica, no són capaços de capturar i mantenir el contingut d'informació de senyals que canvien ràpidament, com ara mesures eòliques i elèctriques.

Abstract

The energy sector has undergone drastic changes and critical revolutions in the last few decades. Renewable energy sources have grown significantly, now representing a sizeable share of the energy production mix. Wind energy has seen increasing rate of adoptions, being one of the more convenient and sustainable mean of producing energy. Research and innovation have helped greatly in driving down production and operation costs of wind energy, yet important challenges still remain open. This thesis addresses predictive maintenance and monitoring of wind turbines, aiming to present predictive frameworks designed with the necessities of the industry in mind. More concretely: **interpretability, scalability, modularity** and **reliability** of the predictions are the objectives —together with **limited data requirements**— of this project. Of all the available data at the disposal of wind turbine operators, SCADA is the principal source of information utilized in this research, due to its wide availability and low cost. Ensemble models played an important role in the development of the presented predictive frameworks thanks to their modular nature which allows to combine very diverse algorithms and data types. Important insights gained from these experiments are the beneficial effect of combining multiple and diverse sources of data —for example SCADA and alarms logs—, the easiness of combining different algorithms and indicators, and the noticeable gain in predicting performance that it can provide. Finally, given the central role that SCADA data plays in this thesis, but also in the wind energy industry, a detailed analysis of the limitations and shortcomings of SCADA data is presented. In particular, the effect of data aggregation —a common practice in the wind industry— is determined developing a methodological framework that has been used to study high–frequency SCADA data. This lead to the conclusion that typical aggregation periods, i.e. 5–10 minutes that are the standard in wind energy industry are not able to capture and maintain the information content of fast–changing signals, such as wind and electrical measurements.

Keywords: Wind Energy; Predictive Maintenance; Machine Learning; Deep Learning; Ensemble Learning; SCADA data limitations

Aknowledgments

Studying is hard, pushing the boundaries of knowledge in a given research field is even harder. Without a solid and reliable group of collaborators it is not possible to achieve remarkable results and discover new things. I make no exception, while this thesis sees me as the author I owe a large part of my results to all those people that have assisted, supported, inspired, and bear with me during all these years.

Starting from Jordi Cusidò and Juan Cardenas that have been my tutors along this journey. Their guidance has been fundamental, it has kept me on the right path when I felt lost. Then, I would like to mention the whole team at SMARTIVE. They have the merit of developing the infrastructure that allowed me to quickly and easily accomplish my researches, without them I would have had to built way more things for scratch and dedicate less time to my researches. A special thanks goes to Olga Porro, who shared with me the struggles of a PhD and helped me pushing forward with publications.

The UPC mathematics group, in particular in the figure of Yolanda Vidal has been an important support in my researches. They have helped me seeing difficult problems from a different perspective and enriched my understanding of applied mathematics and statistics. Some of the interesting idea that I have presented in this thesis are the fruit of our collaboration and of the constant discussion we have held along all these years.

I would also like to thank Julia Walgern and Christian Broer for coordinating and making my stay at Fraunhofer IWES a fruitful and fulfilling experience, despite of it being in the middle of a global pandemic. The whole team at IWES deserves a mention, as they made me feel at home. Above all I would like to acknowledge Timo Lichtenstein and Karoline Pelka, who have directly helped me in my researches providing very challenging and interesting questions on the limitations of SCADA data, the fruits of which you can read in this thesis.

All these great experiences would have not been possible without the financial support and the network that EIT Innoenergy and AGAUR have granted me. I am grateful of having received the privilege to obtain the necessary resources to freely study and advance knowledge in the renewable energy field. These opportunities have given me the possibility to meet and know many inspiring and friendly people. The interesting conversation, the human realtionships, and the great memories I have collected along these years are precious treasures I will always bring with me.

Creating tools or new knowledge to share with other people is the aspect that I appreciate the most about science. Without a cooperative and open community advancements would be much slower. I have made extensive use of open source libraries during my researches, and I have tried to give it back sharing my results through open alternatives. One person I would like to thank is Alberto Diaz from UPM university for kindly designing, sharing, and allowing me to use the LaTeX template that has been used to prepare this document. Without I would have had to spend much more time dealing with formatting and trivial aspects of the document and focus less on the contents.

During these years I had the opportunity to meet many interesting people all along Europe. I have discovered new passions, such as climbing (a shout-out to the new friends I have made at Climbat Barcelona), had interesting conversation, and collected wonderful memories. Living away from home also helped me strengthen and understand the importance of a solid group of friends, as the ones that are with me since my high-school times.

Last but surely not least, I would like to thank deeply my family. My mother Maria and my father Antonio have always supported and believed in me. They have visited me everywhere I have gone, and gave me the means to dedicate my time, resources, and efforts to research. I can not thank them enough, I just hope I have made them proud.

I could mention and tell many other unforgettable stories on this long but fulfilling journey, but hundreds of pages would not be enough. In the interest of future readers I will stop here, but truly believe me when I say that this work if the fruit of a great network of people of which I am only the representative.

—*”Science is magic that works”* **Kurt Vonnegut**

Contents

Contents i

List of Figures iii

List of Tables v

Abbreviations vi

1 Introduction 1
- 1.1 Motivation 3
- 1.2 Scope & Objectives 4
- 1.3 Research dissemination 5
- 1.4 Outline 7

2 State of the Arts 8
- 2.1 Data 8
- 2.2 Models 11

3 Deep Learning and Alarms Information Fusion 20
- 3.1 Information fusion 20
- 3.2 Information Fusion via SCADA Alarms and Autoencoder 20

4 Ensemble Learning Frameworks for SCADA Data 37
- 4.1 Ensemble Learning 37
- 4.2 Ensemble Learning for Wind Turbines Main Bearing 38
- 4.3 Improved Ensemble Learning for Wind Turbine Main Bearing 60

5 Information Loss and SCADA Limitations 78
- 5.1 Limitations of SCADA data 78
- 5.2 Quantification of the Information Loss in SCADA data 79

6 Conclusions 111
- 6.1 Summary 111
- 6.2 Suggestions for Future Works 114

Bibliography 116

A Appendix 130

A.1 Poster WindEurope 2018, Hamburg . . . 131

A.2 Posters WindEurope 2019, Bilbao . . . 132

A.3 Abstract WESC Hannover 2021 . . . 135

List of Figures

3.1 Alarm and autoencoder hybrid predictive system pipeline. 27

3.2 Receiving Operator Curve (ROC) of the various windfarms. 30

3.3 ROC curve of the entire dataset, using a unique threshold value for all the windfarms. 31

3.4 Comparison of the results obtained by the three implemented algorithms hybridized with alarms information. 32

3.5 Characteristic curve that relates the active power and the temperature difference on the two sides of the generator bearing. Distribution of the probability density function of the temperature difference across the generator bearing. 33

3.6 Relation between active power and standard deviation of the generator stator temperature and distribution of the probability density function of generator stator standard deviation 33

4.1 Timeseries profiles of the main bearing temperature of a faulty turbine and the average of the wind-farm. 42

4.2 Relation between main bearing temperature and wind speed. Probability density plot of the main bearing temperature of a faulty turbine and the average of the wind-farm. 42

4.3 Diagram of the predictive maintenance solution. 47

4.4 The rolling window train/test scheme used for normality models. The rolling window train/test scheme used for mean and anomaly indicators. 48

4.5 Boxplot of the main bearing temperature. The median is represented by the red line and the mean corresponds to the triangle. 48

4.6 Normality indicator, RMSE by turbine. Timeseries comparison of predicted versus measured value. . . 49

4.7 Anomaly Indicator plots: percentage of anomalous versus total number of points.(B) 3D plot showing normal (blue) versus anomalous (red) points. 50

4.8 Composed indicator calculation scheme and decision threshold setting. 51

4.9 Relation between KPIs and decision threshold value by wind-farm and indicator. 52

4.10 Performance comparison of individual and composed indicators for Windfarm 1 and Windfarm 2. . . 53

4.11 Scatter-plot of each pair combination of basic indicator. 54

4.12 Correlation matrix of the base indicators. 54

4.13 Heatmap of the combined main bearing health status indicator for wind farm 1. Failures are represented by a yellow star. 55

4.14 Heatmap of the combined main bearing health status indicator for wind farm 2. Failures are represented by a yellow star. 55

4.15 ANN proposed architecture. 66

4.16 Scatter plot showing the isolation forest indicator vs. the neural network indicator independently for each WT. 72
4.17 Main bearing health according to normal behavior model. 72
4.18 Main bearing health according to anomaly detection. 72
4.19 Main bearing health according to the proposed ensemble method. 73
4.20 Saturation mask used to determine whether a turbine should be reviewed or not. 73
4.21 Decision threshold (DT) value vs. the different model indicators for the specificity, accuracy, recall, precision, and F1 score metrics. 74
4.22 Performance metrics comparison for three different decision threshold values. 75

5.1 Illustration of temporal aggregation of SCADA operating data at various resolutions 83
5.2 Illustration of the descriptive statistics of wind speed, active power generation, generator speed, temperature of gearbox shaft bearing 1, and voltage of phase A at different levels of temporal aggregation for wind turbine 5. 87
5.3 Calculated p-values of conducted KS tests for all signals at different levels of temporal aggregation . . . 90
5.4 Exemplary illustration of the information loss of the gearbox bearing temperature for a temporal aggregation of 600 s and a raw signal in 1 s resolution. 92
5.5 Information loss results for all the available signals and various aggregation ranges. 93
5.6 Information loss of three exemplary temperature signals for selected aggregation resolutions. 96
5.7 Information loss results subdivided into all wind turbines for the temperatures of Figure 6. 96
5.8 Information loss results for all turbines partitioned by months. 97
5.9 Boxplots representing wind speed and ambient temperature of an exemplary turbine along the year . . 98
5.10 Heat map of the mean average of the aggregation error value per wind speed bin. 100
5.11 Line plots of the range of variation of information loss for the available signals over wind bins of 0.5 m/s width. 101
5.12 Relation between temporal resolution of the aggregated signal and percentage of data below a given error thresholds. 106

A.1 Illustration of the information loss L for a temporal aggregation of 600 s 136

List of Tables

1 List of Abbreviations . . . vi

2.1 Comparison of advantages and limitations of physical and data–based models . . . 11
2.2 Comparison of advantages and limitations of physical and data–based models . . . 14
2.3 Comparison of the discussed predicting models . . . 19

3.1 Sample of the Supervisory Control and Data Acquisition (SCADA) dataset. . . . 24
3.2 Alarm sample . . . 25
3.3 Work order sample . . . 25
3.4 Results obtained using the autoencoder. . . . 29
3.5 Results obtained using an alarm-based predictor. . . . 29
3.6 Key Performance Indicators (KPIs) results summary for all the available windfarms (WF), sorted by turbine manufacturer (Brand) and location . . . 30
3.7 Normality model results. . . . 31
3.8 Principal maintenance intervention occurred during the testing phase. . . . 32

4.1 Sample of SCADA data . . . 46
4.2 Main Bearing failures work order logs . . . 51
4.3 Comparison of the results of individual and combined indicator for a threshold value of 0.95 . . . 53
4.4 Sample of SCADA data for a limited selection of turbines and signals. . . . 64
4.5 Sample of work order data. . . . 64
4.6 Selected SCADA variables to develop the normal behavior model. . . . 65
4.7 Selected SCADA variables used to develop the anomaly detection model . . . 68
4.8 Main bearing failures reported in the work order logs . . . 70
4.9 Confusion matrix of the ensemble method results with respect to the decision threshold (DT) . . . 74
4.10 Metrics comparison among the single and ensemble methods with respect to the decision threshold (DT) 75

5.1 List of SCADA signals analyzed in this study . . . 85
5.2 Illustration of the temporal aggregation scheme . . . 86
5.3 Signals sorted into bins of pmin. . . . 91
5.4 List of the available signals and corresponding interquartile range values used to normalize results for comparability. . . . 94

Abbreviations

The following abbreviations are used in this thesis:

CBM	Condition-Based Maintenance
CMS	Control Monitoring Systems
csv	Comma Separated Value
EAWE	European Academy of Wind Energy
IEA	International Energy Agency
LCOE	Levelized Cost of Energy
NBM	Normal Behavior Model
O&M	Operation and Maintenance
RUL	Remaining Useful Lifetime
SCADA	Supervisory Control And Data Acquisition
SQL	Structured Query Language
WF	Wind Farm
WT	Wind Turbine
AE	Autoencoder
ANN	Artificial Neural Network
DL	Deep Learning
EE	Elliptical Envelope
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
IF	Isolation Forest
ML	Machine Learning
OCSVM	One Class Support Vector Machine
ReLU	Rectified Linear Unit
RBM	Restricted Boltzmann Machine
SOM	Self-Organizing Maps
SVM	Support Vector Machine
CM	Confusion Matrix
DT	Decision Threshold
FN	False Negative
FP	False Positive
IQR	Interquartile range
KPI	Key Performance Indicator
KS	Kolmogorov-Smirnov
RMSE	Root Mean Squared Error
ROC	Receiving Operator Curve
TN	True Negative
TP	True Positive

Table 1: List of Abbreviations

1. Introduction

The large amount of greenhouse gasses emission caused by human activities, and their noxious effect on the Earth climate have reached a point where actions are required. This reflects in a clear stir in Government policies and people's sensibility towards more sustainable ways of producing energy and efforts to decarbonize the system. For example, the European Union with the 2030 Climate Target Plan has set a goal to reduce its carbon emission to at least 55% below 1990 levels by 2030, and by 2050 to be climate neutral [1]. Similar targets are contemplated in international pledges, such as the historical Paris Agreement, or COP21 in which 190 countries have agreed on a strategy to limit the effect of climate change and avoid irreversible consequences [2].

The energy sector, inclusive of heat and transport, is estimated to produce around three quarters of the worldwide carbon emissions [3]. Moreover, the majority of productive processes depends on electricity. Thus, decarbonization of electricity production will lead to beneficial cascade effects along the whole production chain. To address the need for cleaner electricity, installation of renewable energy sources has soared. The International Energy Agency (IEA) reports that in 2020 renewable capacity has grown by 45%, reaching approximately 280 GW, and in 2021–2022 it is estimated to represent around 90% of new power capacity [1].

In response to these policies energy producers are heavily modifying their production mix including renewable energy sources, mostly adding solar photo–voltaic and wind. Large investments have been done in the innovation and improvement of existing technologies leading to rapid advancements. For instance, the cost of producing one MWh of wind energy has sharply dropped by 44–78% by their peaks of 2007–2010, reaching for the most competitive onshore windfarm a

[1] **Source:** *"Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions"* available at: https://op.europa.eu/en/publication-detail/-/publication/b828d165-1c22-11ea-8c1f-01aa75ed71a1

[2] **Source:** *"UNFCC Paris Agreement 2015"* available at: https://unfccc.int/sites/default/files/english_paris_agreement.pdf

[3] **Source:** *"CO_2 and Greenhouse Gas Emissions"* by Hannah Ritchie and Max Roser, available at: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

value of 0.030 USD/KWh without receiving subsidies [4]. Continuous improvements in the whole energy value chain are pivotal to maintain strong the growth of renewable energy adoption.

The new frontier in wind energy are offshore installations. While open–sea conditions pose a significant challenge for what concerns the logistics, they also guarantees higher reliability and better wind conditions. The energy production of a wind turbine is heavily affected by the quality of the wind resource. Obstacles such as: trees, houses, and hilltops disturb the airflow, generating turbulence that decreases energy production. In the open–sea, no such obstacles are present. Moreover, onshore installations are often restricted by their proximity to human activities. While widely considered as a safe technology —especially when compared to fossil fuels— wind turbines still suffer from the so called *"not in my backyard"* effect [2]. Some local population have shown tepid acceptance of these installations, and in some cases they have actively opposed them.

Improvements in the reliability, control, and monitoring of wind turbines are at the base of the large drop in the price of wind energy. Namely, it was reported that wind energy levelized cost of energy (LCOE) fell 39% between 2010 and 2019 [5]. Most energy companies are investing in monitoring systems that allow to anticipate failures by analyzing sensors data. In recent years the interest of researchers in the application of data analysis techniques to monitor turbines has greatly increased, thanks to the vast amount of data available [3].

Life expectancy of wind turbines is commonly estimated around 20 years, and on average one week of downtime per year is required due to maintenance [4]. This is particularly relevant for those turbines that have been installed in the 1990s and early 2000s that are approaching the end of their lifetime. Windfarm operators have adopted a wide range of measures to extend the operative time of their assets, as mentioned in Ref. [5].

In the past, turbine maintenance adopted a mostly reactive approach, problems in turbine components were addressed when they clearly manifested themselves and corrections were unavoidable. A common practice within windfarm maintainers is to perform inspection based on a regular schedule. Waiting for failure does not allow to optimize maintenance costs and can create serious logistic problems in the management of a windfarm. Scheduled maintenance is also not ideal, as failures in turbines can escalate quickly slipping through periodic maintenance and leading to unforeseen problems. Continuous monitoring is the pro–active solution to this problem. Monitoring can be partially automated thanks to the availability of data and advancements in the fields

[4] **Source:** *"The Power to Change: Solar and Wind Cost Reduction Potential to 2025"* available at: https://www.irena.org/publications/2016/Jun/The-Power-to-Change-Solar-and-Wind-Cost-Reduction-Potential-to-2025

[5] **Source:** *"Renewable power generation costs in 2019"* available at: https://www.irena.org/publications/2020/Jun/Renewable-Power-Costs-in-2019

of signal processing and machine learning, making it a valuable alternative to reactive approaches.

Data sources available to study wind turbines' behavior include dedicated sensors used to record vibrations and acoustic emissions in mechanical components, such as the gearbox and bearings of the turbine transmission [6]–[9]. For electrical components currents signatures can be analyzed [10]. All these options are particularly expensive as these sensors are not part of the standard equipment of wind turbines. Moreover, installing additional sensors poses a logistic challenge as operations need to be halted.

A valuable alternative is the utilization of the *Supervisory Control and Data Acquisition System* (SCADA) that is a network of sensors monitoring the status of the turbine. SCADA was originally designed to provide operators with a tool to control the correct operation of the turbine. Clearly, it was not specifically designed to assess and predict the status of the individual components. In fact, SCADA is characterized by low frequency —typically one record for every 10 minutes— and a focus on operating parameters such as power, speed, and temperature [11]. Nonetheless, SCADA is a compelling alternative to costlier monitoring sensors as it does not require installation of additional equipment and it is ready–available to windfarm operators. Due to these characteristics SCADA analysis is becoming popular both in the academia and industrial research.

1.1 Motivation

The objective of this thesis is to analyze and implement predictive maintenance strategies for wind turbines. The focus is on the **adoptability** of the algorithms in the industry. Providing trustworthy solutions is a priority, as well as improving reliability, and automatize the task of monitoring turbines. Ultimately, improvements in fault detection are expected to decrease the cost of wind energy. Thus, boosting the rate of adoption of this technology in the global energy mix.

The literature related to wind turbine maintenance is rich of complex attempts to anticipate failures, ranging from signal processing analyses, to physical simulations, and machine learning algorithms [12], [13]. **Scalability** of solutions is often neglected by researchers, likely due to lack of large scale datasets which include multiple windfarms and different manufacturers. Most researches are developed for individual windfarms or using laboratory simulations. Rarely algorithms are tested on multiple sites characterized by various turbine technologies and different environment conditions. This is a crucial shortcoming of the literature that this thesis aims to address.

All the solutions presented in this work are tested on multiple sites, characterized by diverse environmental conditions and, when possible, solutions are also evaluated on datasets that include different turbine manufacturers.

1.2 Scope & Objectives

Explicit physical modelling of turbine components is avoided in this research. While detailed physical models are useful to determine cause–effect relationships, and understand the ways subcomponents interacts, they are also very challenging to formulate. They require precise information, not always available to turbine owners, and adapting an existing physical model to a new turbine brand or technology is not trivial. Ultimately, in an industrial scenario physical models do not appear as a feasible, nor cost–effective option. Model–free approaches based on data driven models provide a simpler solution to implement; when compared to physical models, data requirements are far less demanding.

The predictive models that are implemented in the wind energy field are designed for an expert audience, typically composed by engineers and technicians with decades of experience in operating and maintaining turbines. It is important that the output of these models is interpretable otherwise corrective actions are not likely to be taken.

Another identified challenge is the heterogeneous nature of wind turbines fleet. It is common for an energy utility to own multiple windfarm's sites in which a wide variety of turbines is available. The different size of turbines and their design can lead to diverse operating conditions and propensity to specific form of failure. Thus, the optimal predictive strategy should be sufficiently general to work with various technologies and components.

The following objectives are set for this research with the goal to develop solutions that can be fruitfully adopted by the industry:

Interpretability

- To be able to explain the decision of the predictive algorithms. Justify the obtained results. Provide support materials that can convince windfarm operators to turn predictions into corrective actions.

Scalability

- To be able to design predictive strategies applicable not only to specific turbines models, but ideally to a wide variety of brands and manufacturers.
- To design solutions that can be easily scaled from few turbines to entire fleets composed by multiple windfarms that can be located and operated in diverse environments and climates.

Modularity

- To define a predictive strategy that can be easily extended, including new algorithms that are able to capture a wide range of failure patterns.

Reliability of predictions

- To achieve good predictive performances; predictions must be trustworthy and should not lead to a large number of unnecessary check–ups or unforeseen failures.
- To provide timely predictions. The warning time preceding a failure should be sufficient for a windfarm owner to schedule maintenance minimizing logistics costs.

Limited data requirements

- To develop solutions that do not require large economic efforts by turbine owners.
- To develop algorithms favoring utilization of already existing data —i.e. SCADA data, alarms logs, etc.

1.3 Research dissemination

The work presented in this research has resulted in the following scientific contributions:

Articles in peer-reviewed international journals

1. **BERETTA, M.; CÁRDENAS, J.J.; KOCH, C.; CUSIDÓ, J.** Wind Fleet Generator Fault Detection via SCADA Alarms and Autoencoders. Appl. Sci. 2020, 10, 8649. https://doi.org/10.3390/app10238649

2. **BERETTA, M.; JULIAN, A.; SEPULVEDA, J.; CUSIDÓ, J.; PORRO, O.** An Ensemble Learning Solution for Predictive Maintenance of Wind Turbines Main Bearing. Sensors 2021, 21, 1512. https://doi.org/10.3390/s21041512

3. **BERETTA, M.; VIDAL, Y.; SEPULVEDA, J.; PORRO, O.; CUSIDÓ, J.** Improved Ensemble Learning for Wind Turbine Main Bearing Fault Diagnosis. Appl. Sci. 2021, 11, 7523. https://doi.org/10.3390/app11167523

4. **BERETTA, M.; PELKA, K.; CUSIDÓ, J.; LICHTENSTEIN, T.** Quantification of the Information Loss Resulting from Temporal Aggregation of Wind Turbine Operating Data. Appl. Sci. 2021, 11, 8065. https://doi.org/10.3390/app11178065

Presentations in international conferences

- **BERETTA, M.; PELKA, K.; LICHTENSTEIN, T.** Quantification of the Information Loss Resulting from Temporal Aggregation of WInd Turbine Operating Data, Mini-Symposium: Data-driven technologies for O&M cost reduction. Wind Energy Science Conference 2021, (WESC2021). Hannover, Germany. https://www.wesc2021.org/fileadmin/wesc2021/themes/7/BoA_-_Theme_07.pdf

Posters in international conferences

- **BERETTA, M.; CARDENAS, J.J.; BLANCO, A.; JARAMILLO, B.** Health Estimation and Failure Prediction of Wind Turbines Components Based on Correlation Changes Among Significant Variables from SCADA data. WindEurope Hamburg 2018.
- **CARDENAS, J.J.; BERETTA, M.; CUSIDÒ, J.** Your Prediction Algorithm needs the right Threshold. WindEurope 2019, Bilbao.
- **BERETTA, M.; CARDENAS, J.J.; CUSIDÒ, J.** How Ensembling can Boost your Classifier Performances. WindEurope 2019, Bilbao.
- **CARDENAS, J.; BERETTA, M.; CUSIDÒ, J.; AUER, G.; IRIARTE, E.** Turbine's Advanced Life Extension by means of Artificial Intelligence. WindEurope 2019. *Awarded first prize in O&M category* https://windeurope.org/confex2019/networking/poster-awards/

1.4 Outline

This document is articulated in six different chapters containing information on the wind energy industry, its current status and challenges yet to be solved. The first chapter provided an overall introduction picturing the state of wind energy industry, its future prospects and its shortcomings. Here is provided a brief description of the remaining chapters:

- **Chapter** 2 depicts the *State of the Arts* of predictive maintenance in the wind energy industry. The available data source are listed, their pros and cons analyzed. Most typical predictive strategies are explained and a rich selection of articles is provided to the reader.
- **Chapter** 3 presents the first article that was published within this research project. The benefits of mixing diverse sources of information, namely SCADA and alarm logs data, are presented. A reliable predictive strategy, based on deep autoencoders and a ranking system used to combine the two information sources, is applied on a large dataset composed of multiple windfarms and different turbine manufacturers.
- **Chapter** 4 further develops the idea of information fusion by presenting an *Ensemble Learning* framework to combine different SCADA–based indicators. Two published works are included in this chapter. The first one focuses on mixing interpretable indicators; the second one puts the emphasis on improving predictive performances partially renouncing to interpretability of the results and exploring the combination of more complex algorithms.
- **Chapter** 5 tackles the limitations of SCADA data. More specifically, it aims to shed some lights on the relation that exists between the acquisition frequency and information content of SCADA signals. A clear framework to answer questions related to the information content of SCADA data is provided. Indications on what is the recommended aggregation frequency of signals are discussed.
- **Chapter** 6 concludes the document, summarizing the most notable contributions of this research to the wind energy field. An outlook of future works is proposed.

Finally, posters and abstracts submitted to international scientific and industry conferences are included in the Appendix A of this document. The posters encapsulate the preliminary studies that characterized this research, dealing mostly with supervised learning approaches based on failure classifiers and their optimization. The abstract presented at WESC conference in Hannover, instead is the proof of concept of the work that is further developed in **Chapter** 5.

2. State of the Arts

In this chapter the research context and the key concepts to frame predictive maintenance in wind turbines are provided. Specifically the different available data sources are analyzed discussing their advantages, disadvantages, and limitations. Then, the predictive approaches that have been presented in the literature are studied highlighting pros and cons of each technique.

2.1 Data

When approaching the problem of design predictive maintenance strategies for wind turbines a wide array of data sources is available. They can be differentiated by the nature of the signal that are measured (e.g. vibration, currents, acoustic emissions, etc.), but also based on the frequency of these signals. In the following, various sources of information are described; their strengths and weaknesses discussed, and the possible use cases presented.

High frequency data

A large portion of turbine monitoring is based on information–rich data, i.e. data having sampling frequency in the order of the kHz. Mechanical parts of the drive–train can be instrumented with vibrations and acoustics sensors. Often, damages in mechanical components such as bearings and rotating axes can be detected from the insurgence of anomalous vibrations and noise, generated by the interference between the component's parts [9]. Signal processing provides a vast array of tools to analyze vibrations, various kind of signal transformation can be useful to extract information. Both time–domain and frequency–domain tools can be used. Hilbert–Huang transform, for example is useful to isolate information and demodulate noisy signals [14]. Ref. [15] explains how to extract frequencies related to the gearbox drive–train by combining Hilbert–Huang transform and a finite impulse response (FIR) differentiator, obtaining very expressive failure features. Another demodulation tool is the Fast Fourier Trasform (FFT) [16]. Bearing monitoring is commonly done through envelope analysis, one of the main challenges of this approach consists in separating information from noise [17]. Gear monitoring can be based on Cepstrum, which provide information over entire families of harmonics [18], [19]. Other approaches based on statistical analysis [20] and machine learning algorithms have been proposed [21], [22]. More recently, the use of neural networks such as convolutional and recurrent nets have become more popular [23]–[25]. Artificial neural networks (ANN), for example have been used in Ref. [26] to detect various damage modes from vibrations' data of wind turbines. An alternative to vibrations and acoustics signals are current signatures; the advantage of this solution is the reduced cost and non–intrusive nature [27]. In the case of electrical components current leakages can be checked, as some typical failure patterns consists in the partial loss of isolation that leads to parasite currents [10], [28], [29]. Important drawback of high frequency data is the necessity of installing dedicated sensors, which are not part of the standard instrumentation of wind turbines. These sensors are expensive and during their installation they often require interruption of operations. Finally, most of this data sources are considered intrusive, as the installation of additional instrumentation might damage the component being monitored [13].

Oil Analyses

The health of the drive–train —and more specifically of the gearbox— can be assessed through the analysis of the lubricating oil circulating in the system. Turbines can be routinely checked, typically once or twice per year, by taking samples of the lubricant and classify and quantify the chemical compounds contained in it. The presence of large amount of iron and other metals are indicative of defects and damages that can affect the correct operation of the transmission [30], [31]. In this field, data analysis and in particular clustering and other unsupervised learning tools can be effective to help technicians interpreting the data and detect anomalies. Oil analyses are useful for moving mechanical components such as bearings and gears. For static electrical components such as the converter and various parts of the generator other monitoring strategies must be utilized. Moreover, the low frequency of the sample collections can lead to missed diagnosis of defective components.

Alarms

Typically, turbine operations are monitored with the assistance of automated alarm systems. These are implemented comparing the status of some key sensors with respect to predetermined threshold values. Whenever critical values are trespassed a record in the alarm log is entered, complete with the timestamp and the status code of the event. While not as popular as sensors data, alarms have been successfully utilized in the literature. For example, in Ref. [32] alarms data was used to estimate the remaining useful life of wind turbine components. A characteristic of alarm data is that it has no fixed frequency, as its nature is episodic i.e. an entry in the log is recorded only when the alarm criteria is not respected and not following a regular schedule. Moreover, alarms tends to create clusters or alarm cascades, as the activation of one alarm often causes others to spawn [33]. In this sense, a crucial step in alarm analysis is to isolate and understand the alarm sequence. The importance of this step and recommendations on how to approach alarm analysis are provided in Refs. [34], [35]. While being a somehow underdeveloped source of information, alarms data has been successfully used to monitor systems in other fields of application, by applying data mining techniques [36], [37].

Work order logs

Logs of the maintenance intervention, of routines and extraordinary reviews of turbines are often available to windfarm operators. Common taxonomy frameworks have been proposed by industry and academia joint project, such as Reliawind [38]. Yet, not all windfarm operators adopt the proposed standards resulting in large inconsistencies between events logs, and obstacles to formulate general processes to extract information. Nonetheless, common characteristics can be encountered, as the date and time of the maintenance is always provided as well as a brief description of the actions taken during the intervention and the list of materials that were used. Additional details regarding the subsystem that was reviewed can be provided, and sometimes the explication of the root cause of the event is available. Event logs have been used in the literature to filter and label data, isolate and learn failure patterns [39]. Events cascade can also be investigated to find common patterns preceding failures [34]. Finally, general statistics of the most common failures and critical components can be determined analyzing event logs [40]–[42]. Standards in the format of this information, the transition to fully machine readable records, and an improvement in the quality of work order logs are necessary steps to move towards data–centric turbine monitoring.

SCADA

SCADA data was initially designed to provide information to verify the correct operation of turbines, and not as a mean to assess the health status of individual subsystems. The number of sensors monitored by SCADA can vary between turbine manufacturers, though in general the major components of the turbine are all instrumented. The resolution of SCADA data can vary, but most commonly is of 5–10 minutes; only in rare occasion high frequency SCADA data —which can have a reading every few seconds— is available. Physical quantities such as temperatures, speeds, pressures, and states of the turbine are included in a SCADA dataset. Some valuable characteristics of SCADA data is that it is available and standardized for most turbines, meaning that algorithm for its analysis are more easily transferable from one turbine manufacturers to another. Moreover, being part of the standard instrumentation it does not require additional investments by the windfarm owner. The importance of SCADA data for predictive maintenance and monitoring has greatly increased in the last decade. Refs. [43]–[46] are some of the first attempts to use SCADA data for turbine condition monitoring. The methods to analyze and extract information have greatly improved from the early days. In the literature are available algorithms to assess the health of all major components using diverse approaches based on statistical analyses, machine learning, and deep learning [12], [13], [47]. SCADA benefits from very desirable characteristics: the wide availability, a highly standardized format, and its low cost. Nonetheless, its low frequency has been mentioned as an important limitation that can hinder the capability of correctly modeling the status of a turbine and detect failures [39], [48]. Considering the scarcity of researches addressing the topic of information content (and its loss due to data aggregation) in SCADA data, part of this work is dedicated to tackle this problem and shed some lights on the shortcomings of SCADA.

	High–frequency data	**Oil samples**	**Alarms**	**Word Orders**	**SCADA**
Cost	high	high	low	low	low
Frequency	very high	very low	episodic	episodic	low
Standard Implementation	no	no	yes	yes	yes
Methods of analysis	signal processing; time–domain; frequency–domain	standard chemical analysis; unsupervised ML	text mining; graph analysis	text mining; simple statistics	ML; DL
Format	numerical time series	chemical compounds proportion	categorical data	text data	numerical time series
Collection method	vibrations, acoustics, currents sensors	samples taken from drive train components	logged by the SCADA system	logged by the maintainers	logged by the SCADA system

Table 2.1: Comparison of advantages and limitations of physical and data–based models

2.2 Models

Turbine modeling can be divided depending on the final objective of the analysis into two categories: fault prediction, and fault detection [13]. Fault detection aims to assess the status of a turbine and report an occurring problem such that it can be addressed by the maintenance team, as shown in Refs. [49], [50]. Fault prediction has a more ambitious goal, it aims to determine common patterns that precede faults, and use them to anticipate failures [51], [52]. Another criteria for classify models is their approach, that can be physical or based on data.

Physical models

This approach requires deep knowledge and high level of expertise of wind turbine operation principles. Monitored components are modeled into systems of physical equations able to describe their behavior from a thermodynamic, electrical, or mechanical perspective. These are useful to determine and capture how the various components of turbines work, but they also require a vast amount of information and details, which are not always available to researchers. Sometimes, even turbine owners do not have all the required information as manufacturers do not always share the details of the turbines inner systems. Moreover, adapting a model to a different component is not trivial or even possible. Nonetheless, such models have been presented in various studies in the literature. In Ref. [53] a physical simulation of the loads of a turbine gearbox is proposed, showing that it can determine the effect that varying loads have on the lifespan of the component; It required a dynamic study of the gear conditions, as well as a Finite Element Method analysis. Refs. [54], [55] attempt a different approach by first determining the thermal network that describes gearbox conditions; then machine learning is used with the output of the physical model to improve performances. This kind of approach is particularly attractive, it guarantees the possibility to make use of expert knowledge and provide more interpretable results leveraging the predictive power of machine learning and data. This strategy is mimicked in this research, by using machine learning algorithms to capture well-known failure patterns that are commonly monitored by turbine maintainers and automatize their detection. Overall, physical models are a valid option for better understanding the inner working of turbine components and generating new knowledge about them. In fact, physical models are far more reliable than data–based ones, when cause–effects relationships must be determined. The demanding data requirements, the availability of the necessary design parameters, and the scarce re-usability of the models are the motives why different and more flexible tools are required by the industry.

Data–based models

The main alternative to physical models are data–based models. These have risen in popularity thanks to the great advancements achieved in machine learning and statistical modeling. Only minor assumptions of the systems under analysis are needed, as the physical relations governing the operation of the various components are inferred from the data. The resulting models tends to be more transferable; they might need retraining rather than whole redesign when they are used to model different components. On a first approximation, data–based models can be divided depending on how the predictive problem is framed. The two macro families of machine learning problems are: *Supervised* and *Unsupervised Learning*. The first one requires the availability of labeled data that, in the case of predictive maintenance, can be an indication of the health status of the component under analysis (i.e. faulty or healthy conditions). The second approach does not requires labels as useful information is directly extracted from the data. The focus is primarily on the distribution and hidden structures underlying data, rather than a division based on labels.

Supervised learning, in turns can be divided in two major subcategories: *classification* and *regression*. Classification uses data labels to find the characteristics that best differentiate between classes of categorical data. The learned characteristics can then be used to classify new unlabeled data. Regression aims to predict a continuous output. In other words, the value of one or more target variables is predicted from a set of inputs.

Unsupervised learning applications are typically clustering analyses or anomaly detection. Often data is not equally distributed in the feature space, but rather grouped in various clusters. Analyzing clusters in which data falls can help to determine non-obvious relations and groups in the data. It might be found, for example, that anomalous operating conditions separate from the rest of the data and can thus be isolated via clustering and anomaly detection algorithms.

Both approaches have pros and cons. For example, supervised learning and classification problems require complex pre-processing pipelines for data preparation. Reliable data labels must be assigned, then data must be accurately filtered and often enhanced by generating additional features that can highlight interesting properties of data. Finally, the classification model can be trained and used to make predictions. All these steps are time consuming, assigning labels is probably the most critical one, surely the most time consuming. Work orders and alarms logs can be used to assign labels, but the absence of standard formats, the free-text nature of the data, and in general inconsistencies of information pose a great challenge in automating the procedure, as well as guaranteeing reliable labels.

Regression requires a less complex pipeline, labels are used to filter data rather than define groups. A common challenge in regression task is posed by the choice of the distance metric that is used to quantify the quality of predictions and asses the difference between the predicted and observed values when new data is analyzed. Clustering and anomaly detection are far easier to train, and pre-processing pipelines are simpler, although results analyses, choice of the number of clusters and their interpretation is crucial and not trivial.

	Advantages	Limitations
Physical Models	• provide a clear explanation of the system • can be used to determine cause–effects relations • very interpretable results	• require expert knowledge of the modeled system • require detailed information of the system • necessary data is not always available • adapting models to different technologies is not trivial
Data–based Models	• prior knowledge of the modeled system is desirable, but not mandatory • reusable predictive framework • general and scalable approach • can lead to new insights from the data	• require historical data for training • predictions quality depends on input data and the processing strategy • pre–processing data can be time consuming

Table 2.2: Comparison of advantages and limitations of physical and data–based models

Having described the general framework of machine learning and its different variants it is now convenient to dive deeper into the implementation of these predictive frameworks in the research literature. The main approaches are listed below:

- *Power Curve Modeling*
- *Signal Trending*
- *Normality Behavior Modeling*
- *Anomaly Detection*
- *Fault Classifiers*

Power curve modeling

Power curve modeling can be used as a general assessment of turbine status. Monitoring the power output of turbines greatly helps detecting inefficiencies and under-performances. The power curve relation, which is one of the governing equation in the operation of a wind turbine is used by this method. Due to its central role in wind energy extensive researches have been conducted both by academia and industry to determine the parameters that might influence the relation, as well as studying the different methods that can be used to infer it from the data [56]–[59]. For example, the power curve can be approximated fitting a wide range of polynomial equations and determine the one that best suits the data [56]. Model–free algorithms, such as neural networks, are a valid alternative that does not require specific assumptions of the relations within the data [60]–[62]. From a predictive maintenance perspective, power curves can be a useful tool to determine general problems in turbines. The difference between the expected and measured power output is used as indicator of the performance of the turbine. Large and repeated differences between the expected and real power can be a signal of problems in any major system [63]–[65]. Data requirements are limited, as for the easier models only windspeed and output power are needed, more complex models might require additional information, such as turbine location, environment temperatures, and topology of the windfarm. It is to be remarked that power curves are not a complete tool for turbine predictive maintenance strategies. Anomalies, problems, and inefficiencies can be detected but identifying the root–cause, and plan corrective measures analyzing only the power curve is not feasible.

Signal trending

A valid strategy to detect failures in turbine main components is tracking changes and trend in the most significant signals. This approach will be referred to as *"Signal Trending"*, since the objective is to analyze the signal time series and determine trends over time. Ref. [44] shows that gearbox temperature can be a reliable predictor of incipient failures as clear trends in the data can be isolated. The relation between the binned active power and the generator bearing temperature can be compared with respect to the behavior of the windfarm to monitor turbine status, as Ref. [66] shows. Control chart is a tool that is commonly used in conjunction to signal trending to better track anomalies, as shown in Refs. [64], [67]. Alternatively, thermodynamics and physical models can be used to describe the component behavior, then these models can be used to compare loss of efficiency and increasing temperatures such as in Ref. [68]. This class of methods is very heterogeneous, different relations and approaches have been tested by researchers. A common characteristic to these methods is the use of the windfarm as reference for comparisons. The use of simple relations and statistics makes signal trending a very understandable approach. Results are often deviations from a reference value, the relations are based on common thermodynamics, mechanics, and electrical equations which are well known to technicians. Yet, these relations are not always sufficient to capture all kind of failures, and controlling the influence of external factors (such as the environment temperature) is not trivial. More complex approaches are needed to consider additional effects and improving predictions results.

Normal behavior models

A very popular and effective methodology to assess the status of turbine components are the so-called *Normal Behavior Models* (NBM). They attempt to capture the relation between a group of input variables and one or more target signals, that should be able to determine the status of the component under analysis. Important step in NBM is filtering data to determine a subset of records that can be labeled as *"normal data"*. Events and alarms logs are often used for this task. Then, a group of inputs and a target variable can be chosen and an algorithm is fitted to the data. The goal is to infer the model that describes normal operating conditions of the monitored component, and then use it to track deviations between the expected and observed behavior of the tracked variable. This approach is quite common in the wide predictive maintenance world, not only in wind turbine monitoring. One of the first, and probably most influential, application of normality models to wind turbine monitoring is presented in Ref. [43]. Neural networks are a popular choice thanks to their ability to model complex relations in the data, some examples are provided in Refs. [69]–[71]. Other authors have compared the performances of neural networks with standard machine learning algorithms, and regression models [12], [72] or alternative neural networks such as Extreme Learning Machines [73]. NBM allows to capture the complex relations that can exists between turbine operating parameters, such as temperatures, pressures, etc. and external factors such as windspeed and environment temperature. The improved modeling power comes with the cost of lower interpretability if compared to signal trending methods. Tweaking the models, defining the best hyper-parameters and architecture is not trivial and requires knowledge of data modeling and machine/deep learning.

Anomaly detection models

A characteristic of turbine failures is their relatively low frequency, some components are more affected by malfunctions but overall during the course of a year only few occurrences of a given failure are reported [74]. Multiple turbines failing simultaneously due to the same defect are uncommon, except for failures that affect the whole windfarm such as damages to the substation and the connection to the grid. This observation allows to frame the problem as anomaly detection. Clustering, one-class classification, and in general unsupervised learning methods based on point density are useful to detect anomalies. In the literature Self-Organizing Maps (SOM) have been proposed to extract information on turbine status [75], [76]. Machine learning algorithms adapted for anomaly detection or classification of a single class are thoroughly discussed in Refs. [77], [78]. Signal reconstruction is a class of methods that aims to synthesize a generic representation of the data that can be used to filter out unnecessary information, noise, and reconstruct signals. This approach stands at a middle-point between NBM and density-based clustering techniques. In facts, as for NBM a selection of normal data must be available to construct a model of normal data. In contrast to normality models, no single or multiple target variable is defined, instead the whole data distribution is modelled. Algorithms that can serve this purpose are Autoencoders (AE), Generative Adversarial Networks (GAN) or Restricted Boltzmann Machines (RBM) [79]–[81]. A general pattern to these methods is the goal to represent the data distribution, and then track a similarity metric defined by the distance between the original and reconstructed signal, as shown in Refs. [82]. Clustering can provide useful representations of the data, that can lead to interesting discoveries on turbine operations and failure patterns. Moreover, anomaly detection and clustering do not require a fine-grained pre-processing of data nor labeling of operating conditions. Thus, more generic predictive strategies that can be easily adapted to different components can be developed. Signal reconstruction requires pre-processing similar to normality models. The interpretation of the results of anomaly detection and clustering algorithms is critical and it is not always straightforward.

Fault classifiers

Labeled failure data can be used to train a fault classifier which models failures' patterns and outputs predictions, rather than deviations from a normal behavior or anomaly indicators. This is an advantage over NBM and anomaly detection algorithms whose output often requires threshold that must be checked to raise alarms. Assigning labels can be a very time consuming task, work orders and alarm logs must be analyzed and used. A common standard in the industry for the organization and formatting of this information is not available yet, thus different manufacturers may produce very diverse logs. Data labeling can be hardly generalized, re-utilizing the work done for a certain turbine technology on a new one is not easy. Ref. [51] presented a fault diagnosis and prediction solution that is tailored to generator failures; machine learning algorithms such as support vector machine and artificial neural networks are compared . Similarly fault diagnosis and predictions for feeder, generator, and other components using support vector machines is discussed in Ref. [83]. Classifiers have also proven effective with vibration data, as presented in Ref. [84]. The design of preprocessing pipelines of fault classifiers is a difficult task. Two important steps are outliers filtering and feature selection. Having a clean datasets and poignant features is crucial to achieve good performances. Refs. [85], [86] are two valuable contributions dealing both with outlier filtering and feature selection. First, it is highlighted the risk of filtering valuable failure patterns by blindingly using common statistical filters. Second, a complete overview of common feature selection algorithms is presented and their performances are compared to a manual selection performed by an expert maintainer. Another contribution in the field of feature extraction and data augmentation is presented by Ref. [87] in which a discriminative dictionary learning strategy to improve failure predictions in turbines' bearings is implemented. An important challenge for fault classifiers is data imbalance. Faulty conditions are relatively scarce when compared to normal operations. This poses a problem for classifiers. In facts, data imbalance can lead to sub-optimal definition of the boundary that separates classes of data. Refs. [50], [51], [88] faced this problem and suggested various approaches to address it. Another challenge is the wide range of operating conditions that characterize wind turbines. This can lead to errors in the predictions, a two stage process based on clustering of analogous operating conditions, and then classify failures, had positive outcomes in Ref. [89]. Interpretability of fault classifier results is quite limited, predictions are straightforward as typically the outcome of these classifier is a binary one, i.e. faulty or healthy conditions. The factors that have lead the algorithm to these conclusions though is less intuitive, and it depends heavily on the quality of data labels. This, the limited re-usability of trained models, and the time consuming pre-processing procedure are the major drawbacks of fault classifiers.

	Explainability	Complexity	Limitations
Power curve	high	low	• does not identify root-cause • can only analyze the turbine as a whole
NBM	medium	medium	• status of the system should be representable by key variable • struggles with yaw and pitch systems
Anomaly detection	medium-low	high	• typically requires thresholds to raise alarms • it can detect undesired anomalies
Fault classifier	low	high	• preprocessing can be very demanding • labeling quality is critical
Signal trending	high	low	• does not work well with small windfarms or turbines operating under very different working conditions

Table 2.3: Comparison of the discussed predicting models

3. Deep Learning and Alarms Information Fusion

3.1 Information fusion

Combining different information sources is an old and popular approach in data mining [90]. Yet, in wind turbines' predictive maintenance field, SCADA and alarms data are rarely used in the same model. In the following experiment the combination of SCADA and alarms outperformed models which did not use information fusion.

It is important to remark that information fusion leads to improved performances under the assumption that the different data sources are complementary. As shown in the research paper the correlation between SCADA and alarm data is low, thus positive results can be obtained from their combination. Very correlated indicators would have not lead to vast improvements in the results.

3.2 Wind Fleet Generator Fault Detection via SCADA Alarms and Autoencoders

The first contribution of this research involves the combination of two very different data sources: SCADA numerical time series and the alarm logs of the turbines. The generator system has been analyzed due to its important role in the correct operation of wind turbines and the impact that unexpected failures may have on lifetime costs of these assets. The generator, together with the gearbox was found as one of the main downtime causes of wind turbines in Ref. [91]. Being able to detect failures early is crucial to optimize maintenance, reducing costs, and minimizing downtimes.

The chosen architecture combines an autoencoder and an algorithm that analyzes alarm data. The autoencoder is used to model normal operating conditions of turbines, and then, detect anomalies in time series data. Alarms are first filtered, then a selected group of critical events is monitored and the occurrence of these events is used to create a health ranking for each turbine. Finally, the two information sources are combined into a unique status indicator of the generator conditions.

The result is an interpretable solution to assess generator health status. Both data insights and expert knowledge are incorporated in the model, as the selection of the most relevant alarms to track, highly benefits from the experience of turbines' maintainers. Moreover, this framework can be easily extended, including new indicators or sources of data.

Contributions

The main contribution of this chapter are:

- **Present** an information fusion application which allows to use both SCADA and alarms data.
- **Implement** a reliable and interpretable predictive strategy for turbines' generator health conditions.
- **Provide** an industry–ready solution, tested on multiple windfarms located in different locations and environments.

The results have been obtained from the analysis of a large and diverse sample of turbines. A total of 115 wind turbines produced by four different manufacturers and located in three countries were studied. The size and heterogeneity of the training and test sample is uncommon in the wind predictive maintenance literature, which is often limited to single wind farms and manufacturers.

applied sciences

MDPI

Article

Wind Fleet Generator Fault Detection via SCADA Alarms and Autoencoders

Mattia Beretta [1,2,*], Juan José Cárdenas [2], Cosmin Koch [2] and Jordi Cusidó [2,3,*]

[1] Unitat Transversal de Gestió de l'Àmbit de Camins UTGAC, Universitat Politécnica de Catalunya, 08034 Barcelona, Spain
[2] SMARTIVE S.L., 08204 Sabadell, Spain; juan.cardenas@geniussportsmedia.com (J.J.C.); cosmin.koch@smartwires.com (C.K.)
[3] Enginyeria de Projectes i de la Construcció EPC, Universitat Politécnica de Catalunya, 08028 Barcelona, Spain
* Correspondence: mattia.beretta@smartive.eu (M.B.); jordi.cusido@upc.edu (J.C.)

Received: 27 October 2020; Accepted: 1 December 2020; Published: 3 December 2020

Featured Application: Novel approach to wind fleet generator fault detection using Supervisory Control and Data Acquisition (SCADA) data and alarm logs.

Abstract: A hybrid health monitoring system for wind turbine generators is introduced. The novelty of this research consists in approaching a 115-wind turbine fleet by using the fusion of multiple sources of information. Analog SCADA data is analyzed through an autoencoder which allows to identify anomalous patterns within the input variables. Alarm logs are processed and merged to the anomaly detection output, creating a reliable health estimator of generator conditions. The proposed methodology has been tested on a fleet of 115 wind turbines from four different manufacturers located in various locations around Europe. The solution has been compared with other existing data modeling techniques offering impressive results on the fleet. An accuracy of 82% and a Kappa of 56% were obtained. The detailed methodology is presented using one of the available windfarms, composed of 13 onshore wind turbines rated 2 MW power. The rigorous evaluation of the results, the utilization of real data and the heterogeneity of the dataset prove the validity of the system and its applicability in an online operating scenario.

Keywords: alarms; anomaly detection; autoencoder; fault detection; SCADA data; generator; predictive maintenance; wind turbines; renewable energy

1. Introduction

Wind energy is one of the main enablers of the ongoing renewable energy revolution. It was reported by WindEurope that in 2016, wind energy production overtook coal as the second largest form of power capacity in Europe, right behind natural gas. The strong increasing trend suggests that it is just a matter of time for wind energy to take the lead [1].

Many challenges are yet to be solved to increase wind energy profitability, and operation and maintenance (O&M) in particular has to be improved. It was reported that unexpected breakdowns typically cause 10–15% of production losses, with extreme peaks of 30% [2]. These losses cripple the profit of energy companies, thus it is not surprising to find optimization of O&M through big data, cloud solutions and innovative technologies as one of the top priorities of the industry [3].

Historically, maintenance has been performed via a reactive approach, based on preventive inspections and corrective interventions once failures were acknowledged. New approaches providing predictive maintenance solutions have emerged both in the academia and the industrial scene.

Turbines are commonly equipped with a Supervisory Control and Data Acquisition (SCADA) system, which was initially installed to monitor and operate the system, but lately has been utilized to assess and predict the health status of the turbines as well. SCADA data is recorded by a network of sensors located in the main components of the turbine, the typical sampling frequency is 10 min, making it relatively cheap to collect, transmit and store in a database. All wind fleet operators collect data on their centralized control. SCADA data is collected and stored on Structured Query Language (SQL) databases from SCADA providers or OsiSoft PI system.

Early fault detection can be achieved, as shown by Schlettingen and Santos, by building a model that captures normal operation of the system and by comparing the difference between predicted and measured values of a key variable, to detect anomalies [4]. This approach does not fully take advantage of the high dimensionality of the SCADA dataset and focuses only on the behavior of a single key variable, while component failures are typically complex and can manifest themselves in different failure modes.

The literature is rich with examples based on power curve modeling of wind turbines [5–8]. This approach is based on tracking the relation between wind speed and output power, the function that describes the relation between these two variables can be inferred from operational data and compared to the one provided by the manufacturer, and significant deviations from the theoretical power curve can be hints of problems in the turbine. Different algorithms, as well as the introduction of context variables, have been studied in order to get a reliable picture of the turbine behavior. The main drawback of this approach is its incapacity to determine which component is causing underperformance since the turbine is studied as a whole.

Solutions based on control monitoring systems (CMS) are available and have been studied in the literature [9–11]. These analyses typically use vibration, sound and acceleration measurements to detect anomalies in the behavior of bearings, gearboxes and other mechanical components. The frequency of the data used for these analyses is much higher than the typical SCADA data, thus bringing more information for the detection of failures. That being said, most turbines are not provided with vibration sensors, the installation of these instruments disrupts the operation of the turbine and can cost a windfarm owner thousands of euros per turbine. The authors of Reference [12] presented a thorough analysis of the available monitoring techniques for wind turbine; regarding the CMS, they highlighted as main challenges: financial cost, difficulty of interpretation of the results and not-trivial integration with all the existent monitoring systems, as well as its scalability.

For these reasons, solutions based on the usage of SCADA data can be particularly interesting for owners of old turbines, since no installation of additional sensors or interruption of their operations is needed. Value can be created from the large quantity of unutilized SCADA data stored in their databases.

The rapid growth of the Deep Learning field led many researchers to apply neural networks to solve data challenges. Autoencoders in particular appear to be a good fit for anomaly detection. Autoencoders have been applied in multiple practical applications, such as anomaly detection of seasonal Key Performance Indicators (KPIs) in web application [13], cyber-security monitoring [14] and monitoring of gas turbine conditions [15].

In the wind energy sector, Jiang et al. stacked multiple autoencoders to extract new representations of vibration data in the event of gearbox failures [16]. Successively, they also utilized denoising autoencoders, enriched with temporal information to assess turbine conditions in a laboratory and online scenario [17]. Finally, autoencoders have been successfully used for ice-detection on turbines' blades by Liu et al. [18].

Alarms and events records have been used to determine the remaining useful life of wind turbines [19]. In Reference [20], the time-sequence of the alarms is analyzed to detect relations between the different alarms, determining the causal relationship between the different events and helping to determine the root-cause of failures.

This research aims to explore the capabilities of autoencoders and SCADA alarms as a hybrid fault detection system for wind turbines' generators. While in the literature examples of predictive strategies based only on SCADA data or alarms are present, no holistic approach using both sources of information is present. This paper reports a methodology that takes advantage of both SCADA and alarm logs in the same algorithm.

As a benchmark, other typical anomaly detection algorithms are implemented, and their results are compared with the autoencoder's results. Additionally, the overall methodology is compared to a normality model, one of the most common predictive maintenance approaches available in the literature. Given the practical nature of the project, SCADA and alarm logs of existing windfarms are used. Results are validated using maintenance logs and verifying the concordance between the predictions and the available information.

A key aspect of this investigation is the thorough analysis of real data from a heterogeneous sample of data. The dataset includes four different turbine brands, from seven different windfarms, located in different nations and climates (Spain, United Kingdom and Poland). Moreover, the size of the sample is remarkable, as more than a hundred turbines are studied. These factors are rare in the relevant literature, as most of the time, a single turbine or windfarm is analyzed. All these considerations support the applicability of the approach in real-life scenarios and its ability to generalize results to heterogeneous conditions.

2. Materials and Methods

2.1. Data Description

The source of information used for this research are the SCADA and alarm datasets as inputs to the model, and the maintenance task logs as ground-truth material to evaluate the effectiveness of the methodology. Two years of operation data for more than 100 turbines rated 2 MW and different manufacturers was available. Data has been received directly from the windfarm operator in the form of comma-separated values (csv) and text archives and uploaded in a SQL database.

The dataset was split into a training and test set, maintaining a train/test split ratio of 70–30%. The last 9 months of data have been used as the test dataset, and the remaining data was used for training the algorithms. The utilized data is a real-life dataset of various windfarms operating under common conditions, it is not the results of a simulation. As a consequence, the data required thorough cleaning and pre-processing to get rid of inconsistencies due to sensors' errors and communication malfunctions.

2.1.1. SCADA Dataset

The SCADA dataset contains more than 300 variables as the main systems of the turbine are all monitored (pitch, main shaft bearing, gearbox, generator, etc.). Sampling frequency is 10 min and quantities such as the arithmetic mean, minimum, maximum and standard deviation are computed with the data acquired for this period. The format of the SCADA dataset, as well as the name of the variables and position of the sensors, may vary according to the manufacturer of the turbine. An example of the dataset used in this research is provided in Table 1.

Table 1. Sample of the Supervisory Control and Data Acquisition (SCADA) dataset. Average (avg.) and standard deviation (std.) values are reported.

Timestamp	Wind Speed (avg.) (m/s)	Power (avg.) (kW)	Power (std.) (kW)	Generator Stator Temperature (avg.) (°C)
2018-10-01 00:10:00	4.945	282.8	28.524	64.653
2018-10-01 00:20:00	5.361	331.433	20.253	64.322
2018-10-01 00:30:00	5.01	289.525	47.297	61.16

2.1.2. Alarm Dataset

Alarms are typically triggered whenever an operating parameter, most typically a temperature, exceeds its normal operation range. Table 2 is an example of the information contained in the alarm dataset. The alarm description field contains standardized text data, generated by the control system of the turbines.

Table 2. Alarm dataset sample.

Start time	End time	Turbine	Alarm Description
2017-05-24 10:39:29	2017-05-24 10:40:27	WT05	Gen brushWear Warn
2018-07-05 11:47:29	2018-07-05 11:59:57	WT03	GenRot RpmMonitor Stop
2018-04-13 08:00:58	2018-04-13 12:37:04	WT02	Gear OilFilt Warn (75% clogged)

2.1.3. Work Orders Dataset

All the maintenance tasks that have been carried on in the windfarm, including inspections, regular checks as well as extraordinary interventions, are registered in the work order logs. An example of the available work order logs is provided in Table 3. This information has been used for labeling turbines' SCADA data. Records preceding critical interventions to the turbines have been removed from the training dataset. Work orders have also been used for the prediction evaluation. The information of the work orders is not provided in any form to the predicting algorithm, it is uniquely utilized to process data, assigning labels, and finally, evaluate the predictions, thus being the ground truth for the algorithm.

Table 3. Work order sample.

Start Time	End Time	Turbine	Component	Work Description
2017-02-18 07:52:00	2017-02-19 13:30:00	WT06	Generator bearing	Generator bearings replacement
2017-06-30 10:46:00	2019-06-30 14:03:00	WT08	Blade	Scheduled inspection
2017-08-27 08:50:00	2018-09-03 15:28:00	WT07	Gearbox	Gearbox replacement due to fractured gear tooth

2.2. Autoencoder Anomaly Detection

Anomaly detection via autoencoder is performed providing the network a training dataset composed of normal data, that can be represented as $\{x(1), x(2), ..., x(m)\}$. Autoencoders can be divided into two parts: an encoder and a decoder.

The encoder's goal is to reduce the dimension of the data, mapping data into lower dimensional spaces, reducing the number of neurons in each successive layer, until the bottleneck is reached. The number of layers and neurons in the network is determined by a tradeoff between the compression of the input information and the ability to reconstruct the input sufficiently well. Neurons are activated by an activation function such as the one presented by the following equation [21]:

$$a_i^{(l)} = f\left(\sum_{j=1}^{n} W_{ij}^{(l-1)} a_j^{(l-1)} + b_i^{(1)}\right) \tag{1}$$

where W and b are the weight and bias of the model, and the indexes i and j denote the unit and the layer, respectively. Non-linear activation functions are typically utilized to allow the network to represent non-linear characteristics of the data. In this research, the rectified linear unit (ReLU) function has been used, and is defined as follows:

$$f(x) = \max(0,\ x) \tag{2}$$

The decoder's function is to reconstruct the encoded data at the best of its possibilities. The entire structure, encoder and decoder, is in fact optimized, minimizing the following cost function, presented in Reference [21]:

$$J(W,b) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|x(i) - \hat{x}(i)\right\|^2\right) + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(W_{ji}^{(l)}\right)^2 \tag{3}$$

in which n_l is the number of layers, s_l is the number of units in layer L_l and λ is the regularization parameters that keep a balance between the memorization and generalization capabilities of the network. As the equation shows, a larger and more complex network would be penalized by the factor λ. The first part of the equation defines the difference between the input and output vectors, and thus a priority of the network will be to minimize this difference.

As explained in Reference [22], anomaly detection using autoencoders can be seen as a semi-supervised learning problem. The autoencoder is trained with normal data and learns its representation in a reduced dimensional space. The reconstruction error is utilized as a metric to determine abnormal data. Data that does not fit the representation learned in the training phase results in higher reconstruction error and can be marked as anomalous.

2.3. Methodology

Fusion of multiple sources of information, namely SCADA data anomaly detection and alarm registers, is the core of this research. First, the initial processing of the SCADA data is presented, then the processing of the alarms and the final step of merging the autoencoder and alarms' predictions in unique indicators are discussed separately.

2.3.1. SCADA Data Processing

Of the entire dataset, a subset of six variables is used to model the generator: active and reactive power, temperature of nacelle and generator stator, as well as wind and generator speed. While the dataset was composed of more than 300 variables, just a small selection was kept. Processing all the variables would result in very large computation time and likely lead to overfitting of the data, interpretability of the predictions would also be not trivial since the number of inputs would be very large. The selection of the variables has been done choosing measurements related to the system under evaluation (generator speed, generator stator temperature) as well as context signals that determine the operating status of the turbine (active and reactive power, nacelle temperature and wind speed).

The dataset is split into a training and a test set, the first 70% of the data was used for training and the remaining 30% for test. Data shuffling has been avoided, since the dataset is composed of timeseries and random selection of data could result in information leakage.

Analysis of the maintenance and alarm logs allows to filter out abnormal operating conditions from the training set, as well as remove outliers caused by sensor malfunctions, thus creating a training set composed only by normal operation records. No imputation of missing data was performed. To filter data, pre-processing algorithms [23] are applied. In practice, a range of acceptable values for the input variable of the model is defined and all the data not conforming with this range has been filtered, considered as communication errors.

A crucial part of pre-processing is normalization of data, the training set is used to determine the minimum and maximum value for each input variable, and these values are then stored to be used later on in the test set.

2.3.2. Autoencoder Architecture Selection and Training Process

To determine the optimal architecture (number of layers and neurons, activation function, etc.) of the autoencoder, a grid search approach is used, multiple configurations are tested and the one obtaining the lowest reconstruction error is chosen. Training time and complexity of the network have been considered. A process of trial and error of different configurations is necessary to determine the best structure for the available data; thus, a different dataset could result in a different network structure. The best network layout is a fully connected network composed of six layers, having respectively 7–12–4–12–7 neurons activated by the rectified linear unit (ReLU) function and mean squared error was used to measure the distance between the input and output, and the optimization algorithm is "adam".

Having found the best network layout, its predictions on the training data are created to obtain the distribution of the reconstruction error, which is the difference between the original and the processed data. The assumption that the reconstruction error does not contain systematic errors is verified, analyzing its distribution that resembles a normal distribution. Using this information, it is possible to determine a critical value to identify anomalies. Three standard deviations from the central value are utilized.

2.3.3. Alarms' Processing

Alarms' data is processed by selecting, from all the alarms available in the dataset, the ones that are more relevant for the generator assembly, such as high temperature, overspeed and overload of the generator or its auxiliaries, such as cooling fans. The alarm description field of the dataset was analyzed by keywords, terms such as: "high-temperature", "error", "warning", "over speed", "overload", etc., were searched. In this step, expert knowledge played an important role in excluding from the initial selection those alarms that do not represent truly critical conditions and not simple communication errors.

Once the list of alarms has been defined, it is possible to count how many times any of the selected alarms has occurred during the period under evaluation. In this research, the authors decided not to assign a different weight to the various alarms and simply counted the occurrences. More refined strategies involving rankings of the alarms, as well as detection of patterns or study of the time separating two consecutive alarms, could be implemented in future studies. According to this indicator, turbines having a higher number of alarms should be prioritized for maintenance.

2.3.4. Indicators' Merging Process

The health predictions are made for the entire period of time comprised in the test set and information is aggregated to construct a generator health indicator. Anomalies are summarized to a weekly resolution, by comparing the number of anomalies detected in each turbine with respect to the windfarm. The distribution of anomalies within the windfarm is calculated and turbines lying at a distance superior to two standard deviations from the central value are considered anomalous. This is done because particular external conditions lead the entire windfarm to behave anomalously while not undergoing a real fault in the generator system.

The generator's health indicator is a vector defined in a two-dimensional space. The components of the vector are the processed output of the autoencoder and the counter of key alarms per turbine during the period of the analysis, the module of the vector is calculated as the Euclidean Sum of the two components. A threshold is defined to determine and prioritize the turbines that require maintenance. Alarms' data is used directly in the model, hybridizing and complementing the results of the numerical analysis performed with the autoencoder. The generated status vector considers anomalies in the numerical data and information from the alarm system.

Figure 1 summarizes all the steps of the proposed methodology showing data reception, its storage and preprocessing and the predicting algorithm.

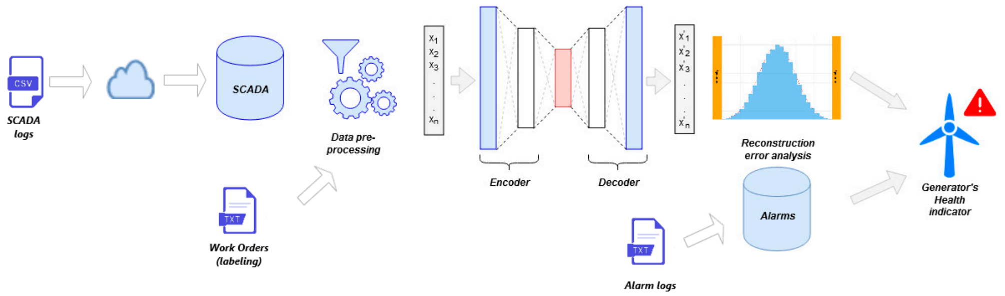


Figure 1. Alarm and autoencoder hybrid predictive system pipeline.

3. Results

The methodology has been proven on a fleet of more than 100 wind turbines, from four different manufacturers, located in very different geographical locations ranging from hot climates, such as south of Spain, to colder ones such as Poland and the United Kingdom. While adjustments were required due to the different variables and characteristics of the turbines, the overall methodology was not modified.

3.1. KPIs Definition

A brief explanation of the indicators utilized for the presentation of the results is provided in this subsection.

In order to assess the prediction power of the predictive models, we have used the confusion matrix (CM) as a basic unit of evaluation. The CM consists of four labels given to each prediction according to its veracity. In summary, these labels are true positives (TP, a failure occurs when a failure was predicted), false positives (FP, no failure when a failure was predicted), true negatives (TN, no failure when no failure was predicted) and false negative (FN, failure when no failure was predicted). Using the count of these basic evaluation units, the main KPIs are calculated.

The main KPIs used in this project are sensitivity, specificity, accuracy, Kappa, precision and F1 score. Sensitivity, Recall is the ratio of predicted events over the total of events:

$$Sensitivity,\ Recall = \frac{TP}{TP + FN} \tag{4}$$

Specificity is the ratio of well-predicted negative events over the total of negative events:

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

Accuracy is the ratio of the total well-predicted observations over the total number of observations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Cohen's Kappa is defined as follows:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{7}$$

where p_0 is the relative observed agreement among raters, which is analogous to accuracy, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. For categories, k, number of items, N, and n_{ki}, the numbers of times the rater i predicted category k, p_e can be calculated as follows:

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2} \tag{8}$$

A low value of K means that there is no agreement among the raters other than what would be expected by chance. A K value close to one is an indication of good performance of the classifier.

Precision is the ratio of predicted events over the total of positive predictions:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

F1 score is defined as the harmonic mean of precision and recall and it is typically used to measure the accuracy of a test:

$$F1 = 2\frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FN + FP)} \tag{10}$$

3.2. Autoencoder and Alarms Results

As one of the main goal of this research is to demonstrate the advantages of merging different sources of information, the results of the autoencoder and an alarm-based predictive system are presented and compared to the numbers obtained using a unique predictor made by the fusion of the two individual methods.

Table 4 presents the results obtained using the autoencoder as a unique predictor of the generator status. It can be seen that various failures are anticipated, but the rate of FPs is quite high, as well as the FNs.

Table 4. Results obtained using the autoencoder. WF stands for Windfarm, TP True Positive, FN False Negative, FP False Positive and TN True Negative.

	Brand	Location	TP	FN	FP	TN	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1
WF1	Vestas	Spain	3	1	4	3	55%	15%	75%	43%	43%	55%
WF2	Vestas	Spain	1	3	2	3	44%	−15%	25%	60%	33%	29%
WF3	Siemens	Poland	1	0	6	11	67%	17%	100%	65%	14%	25%
WF4	Siemens	Poland	3	3	1	8	73%	41%	50%	89%	75%	60%
WF5	Senvion	Poland	4	1	4	4	62%	27%	80%	50%	50%	62%
WF6	Senvion	Poland	3	1	6	12	68%	28%	75%	67%	33%	46%
WF7	Nordex	UK	2	4	7	13	58%	−1%	33%	65%	22%	27%
TOTAL			17	13	30	54	62%	18%	57%	64%	36%	44%

Table 5 shows the results obtained using an alarm-based predictor. The results are not so different from the autoencoder's ones, a slightly higher Kappa is achieved by this method, and one more TP was found, while the FPs rate is almost equal. It is clear that neither of the two techniques, on its own, would be sufficiently reliable in a real-life scenario.

Table 5. Results obtained using an alarm-based predictor. WF stands for Windfarm, TP True Positive, FN False Negative, FP False Positive and TN True Negative.

	Brand	Location	TP	FN	FP	TN	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1
WF1	Vestas	Spain	2	2	1	6	73%	38%	50%	86%	67%	57%
WF2	Vestas	Spain	2	2	1	4	67%	31%	50%	80%	67%	57%
WF3	Siemens	Poland	1	0	17	0	6%	0%	100%	0%	6%	11%
WF4	Siemens	Poland	2	4	2	7	60%	12%	33%	78%	50%	40%
WF5	Senvion	Poland	4	1	0	8	92%	83%	80%	100%	100%	89%
WF6	Senvion	Poland	4	0	6	12	73%	42%	100%	67%	40%	57%
WF7	Nordex	UK	3	3	4	16	73%	28%	50%	80%	43%	46%
TOTAL			18	12	31	53	62%	19%	60%	63%	37%	46%

3.3. Overall Results

As the results of the individual predictors are not sufficiently good, the authors present a hybrid technique that merges the two systems in a more complete predictor, as detailed in Section 2.3. Table 6 presents a summary of the results. The turbines that were obtaining higher values for the health KPIs were reported. Examining the reported turbines and the maintenance log, the results table was done. During the test period, problems such as broken generators, consumed generator brushes or generators bearing damages were encountered.

It can be seen that most of the reported turbines were found to have some problems; moreover, the results across the various windfarms are consistent. The accuracy never gets lower than 70% and the overall Kappa is 56%. The advantages of using a hybrid predictor are clear when its results are compared to the ones of the autoencoder and alarm predictors. The number of TPs increased substantially, and remarkably, the number of FPs was halved. The two components of the composed predictors are complementary, allowing for more accurate and reliable predictions.

Table 6. Key Performance Indicators (KPIs) results summary for all the available windfarms (WF), sorted by turbine manufacturer (Brand) and location. TP stands for True Positive, FN False Negative, FP False Positive and TN True Negative.

	Brand	Location	TP	FN	FP	TN	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1
WF1	Vestas	Spain	4	0	0	7	100%	100%	100%	100%	100%	100%
WF2	Vestas	Spain	3	1	1	4	78%	55%	75%	80%	75%	75%
WF3	Siemens	Poland	1	0	3	14	83%	34%	100%	82%	25%	40%
WF4	Siemens	Poland	4	2	0	9	87%	71%	67%	100%	100%	80%
WF5	Senvion	Poland	5	0	2	6	85%	70%	100%	75%	71%	83%
WF6	Senvion	Poland	3	1	4	14	77%	41%	75%	78%	43%	55%
WF7	Nordex	UK	3	3	4	16	73%	28%	50%	80%	43%	46%
TOTAL			23	7	14	70	82%	56%	77%	83%	62%	69%

The Receiving Operator Curve (ROC) is calculated to represent the predictive power of the proposed methodology and its response to adjustments in the cutoff value to apply to the health status vector. In Figure 2, the ROC curves of the different windfarms are presented. The cutoff values are adjusted for each wind farm to obtain optimal results.

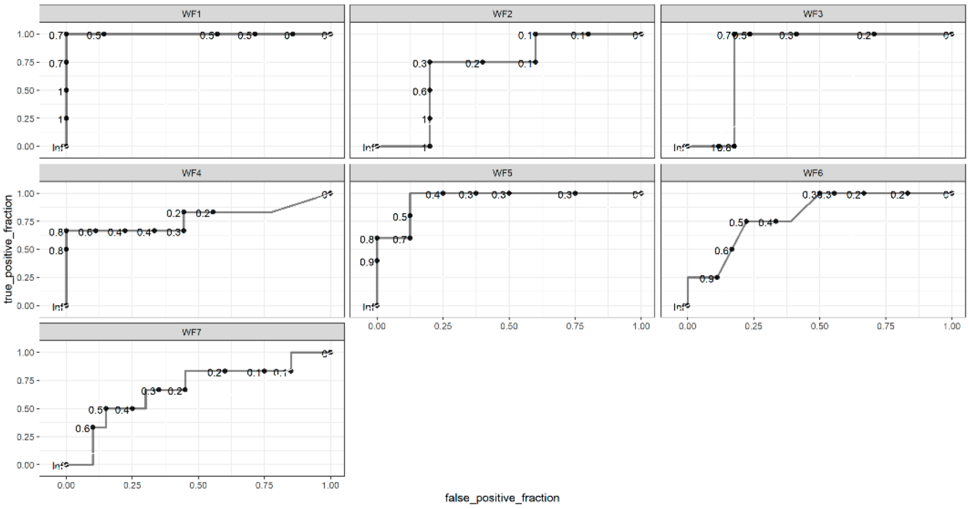


Figure 2. Receiving Operator Curve (ROC) of the various windfarms, the points and the corresponding values represent possible cutoff values to make a decision.

Figure 3 represents the dataset as a whole, without distinction between the different windfarms and simulating the effect of a unique cutoff value. The two dashed line defines the values of the false positive rate and true positive rate that can be obtained by selecting the optimal cutoff value for each windfarm. It can be seen that fixing a unique threshold value yields good results while being a simpler decision strategy, but in applications where the reliability of the prediction is the key objective, the additional complexity provides better outputs.

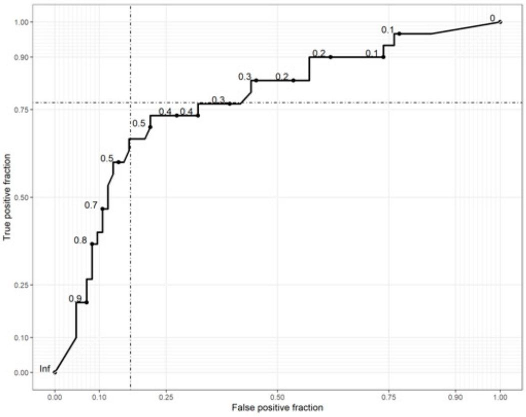


Figure 3. ROC curve of the entire dataset, using a unique threshold value for all the windfarms. The point defined by the intersection of the dashed lines is the result reported in Table 6.

3.4. Normality Model Comparison

An additional validation of the results is presented. A normality model using the same input data is trained and utilized to make health predictions of the generators. Details on how to build a normality model are available in Reference [4]. The value of the generator stator temperature is predicted by a ridge regression model and the prediction error is used as a metric for the generator status. Details on the algorithm can be found in Reference [24], the decision of using this algorithm is dictated by its capacity to deal with multicollinearity in the inputs. The results of the normality model are presented in Table 7**.**

Table 7. Normality model results. WF stands for Windfarm, TP True Positive, FN False Negative, FP False Positive and TN True Negative.

	Brand	Location	TP	FN	FP	TN	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1
WF1	Vestas	Spain	4	0	2	5	82%	65%	100%	71%	67%	80%
WF2	Vestas	Spain	1	3	0	5	67%	27%	25%	100%	100%	40%
WF3	Siemens	Poland	1	0	12	5	33%	4%	100%	29%	8%	14%
WF4	Siemens	Poland	4	2	6	3	47%	0%	67%	33%	40%	50%
WF5	Senvion	Poland	0	5	2	6	46%	−28%	0%	75%	0%	NA
WF6	Senvion	Poland	2	2	5	13	68%	17%	50%	72%	29%	36%
WF7	Nordex	UK	5	1	6	14	73%	41%	83%	70%	45%	59%
TOTAL			17	13	33	51	60%	14%	57%	61%	34%	43%

One can see that while the normality model yields reasonable results, it scores lower overall values for the tracked indicator when compared to the presented methodology. In particular, it should be noted that the number of FPs is more than double the proposed solution and the total number of TPs is lower. The only case in which the normality model performed better is WF7, where two additional TPs are found.

The presented results were obtained using a large sample of real data. The sample is extremely heterogeneous since it represents four different turbine brands, and the windfarms are located in different geographical locations (Poland, Spain and United Kingdom), characterized by very different climates and wind conditions. Such results are rare in the literature, as many algorithms have been tested either in laboratories or in a reduced sample of turbines.

In Section 4, the detailed analysis of windfarm 5 is proposed. This one was chosen since it has a high prevalence of failures of the generator and two predictions were classified as FN, so it is useful to analyze them in detail to determine the reason why the alarms were raised.

4. Discussion

The last 9 months of data available were used as a test set. The performance of the autoencoder as an anomaly detector was compared to other algorithms that have been widely used for anomaly detection tasks. Isolation forest and one-class support vector machine were tested. Details on these algorithms can be found in References [25,26].

The same post-processing methodology was applied to all algorithms. Results are presented in Figure 4. Three risk-areas were identified based on the generator's health indicator value distribution. Table 8 provides the information to assess the accuracy of the predictions, and major component replacements that took place during the testing phase are reported.

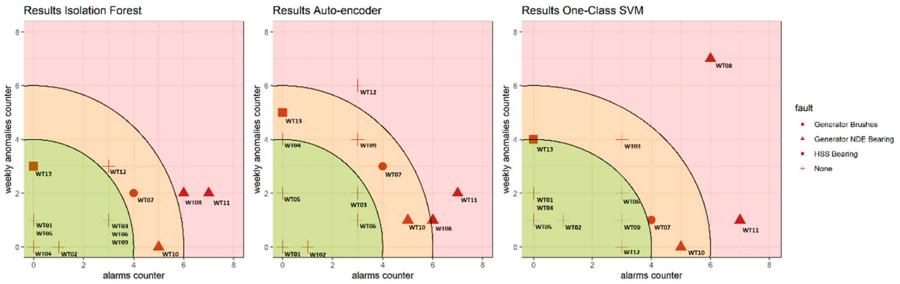


Figure 4. Comparison of the results obtained by the three implemented algorithms hybridized with alarms information. The higher the distance from the origin, the worse the conditions of the generator. Three areas are identified according to the health status: healthy (green), warning (yellow) and danger (red). The shape determines the presence and type of fault occurred.

Table 8. Principal maintenance intervention occurred during the testing phase.

Turbine	Maintenance Description	Component
WT13	Bearing High Speed Shaft replacement	Gearbox-Generator
WT11	Generator brushes replaced	Generator
WT11	Generator bearing Non-Drive End replaced	Generator
WT10	Generator bearing Non-Drive Endreplaced	Generator
WT08	Generator bearing Non-Drive Endreplaced	Generator
WT07	Generator brushes replaced	Generator

All three algorithms, when merged with alarm information, are able to satisfactorily isolate faulty turbines from the rest. Autoencoder is selected as the algorithm of choice to analyze SCADA data, since it is able to better diagnose faulty turbines even in the absence of alarms data, as in the case of turbine WT13. Moreover, the autoencoder better identifies the high-speed shaft-bearing fault, where isolation forest could not separate it sufficiently and one class Support Vector Machine (SVM) positioned it on the frontier between the warning and safe areas, the ability to identify various failure modes holds large relevance in the selection of the algorithm. Analyzing the results of the autoencoder, it can be noticed that most of the turbines in the critical (red) and dangerous (yellow) areas required replacement of the bearings or brushes of the generator. None of the windmills located in the safe (green) area required maintenance.

A detailed study of the data of WT09 and WT12 was done due to their high anomaly count and absence of maintenance intervention. The input variable distributions of all the signals and some other key variables of the generator have been reviewed thoroughly to understand the reason why the autoencoder has found these turbines to be anomalous. The most relevant relationships related with generator failure are presented here and discussed.

In Figure 5, the distribution of the probability density of the temperature difference across the two sides of the generator bearing of turbine WT09 are represented, compared with the mean value of the windfarm and the characteristic curve of this temperature difference with respect to nominal power. It can be observed that the behavior of turbine WT09 is widely different from the rest of the windfarm. These considerations lead us to categorizing this prediction as early fault alert of the generator bearing conditions, rather than a false alarm.

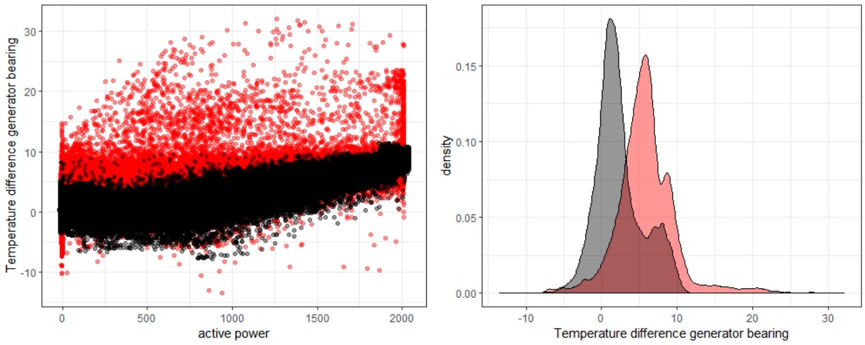


Figure 5. On the left, the characteristic curve that relates the active power and the temperature difference on the two sides of the generator bearing. On the right, the distribution of the probability density function of the temperature difference across the generator bearing. In red, the data belonging to turbine WT09, in black, the mean of the windfarm.

Figure 6 shows that turbine WT12 is characterized by an anomalous distribution of the generator stator temperature, in fact the standard deviation of its recorded values is larger than the value of the windfarm, meaning that the generator of this turbine is subjected to less stable operating conditions. This case can also be considered anomalous and worthy of a technical review of the generator.

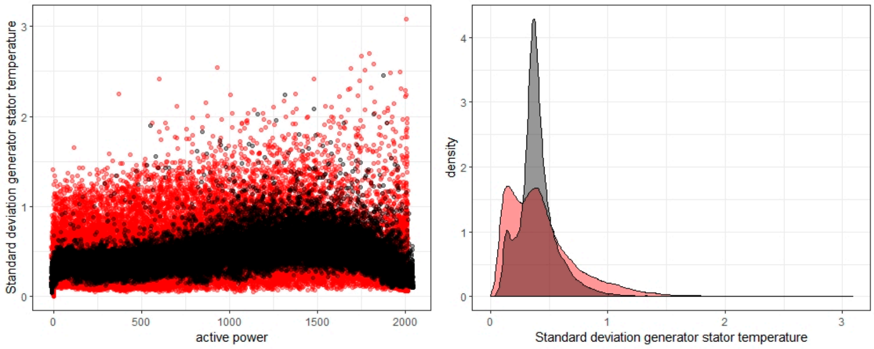


Figure 6. Relation between active power and standard deviation of the generator stator temperature (**left**) and the distribution of the probability density function of generator stator standard deviation (**right**). In red, values of turbine WT12, and in black are the mean values for the entire windfarm.

Merging the information of alarms with anomalies provides a more comprehensive health status of the generator. Looking at the plots, it can be seen that alarms are able to isolate most of the faulty turbines, that being said, there are also cases in which a low number, or no alarms are raised, but nonetheless, the turbine was found to be faulty. WT08 problems are detected mainly by the alarm counter, whereas WT13 is purely diagnosed by the anomaly count, the rest of the faults are found by a mix of the two information sources. Ultimately, merging the information from alarms and SCADA data proved a rewarding strategy able to better separate turbines according to their health status, making use of available and easily accessible data.

5. Conclusions

A hybrid fault detection system based on SCADA alarm logs and an anomaly detection autoencoder were presented and validated on a fleet of more than 100 wind turbines, from four different manufacturers, located in different parts of Europe. Real operating data has been used and most of the raised alarms corresponded to problems related to the generator that required the substitution of the component or some parts of it (bearings, brushes).

A detailed explanation of the most critical windfarm was presented to show how the methodology can be applied in practice and the kind of analyses that were carried out to corroborate the results.

It has been observed that the alarm counter is a valid tool to distinguish faulty turbines from healthy ones. That being said, the alarm counter alone cannot anticipate all failures. The fusion of anomalies and alarms information complements the individual approaches, providing a more reliable system.

All five failures that occurred during the test phase were correctly detected. Of the two "false positive" predictions that were obtained, detailed analyses suggested that they are likely early fault detections, rather than errors. Ultimately, this methodology provides windfarm operators a reliable tool to assess the health of generators and improve operation and maintenance of the turbines.

The results of the autoencoder as an anomaly detector were compared with other common algorithms in the literature, such as isolation forest and one-class support vector machine. The results showed that while the other two algorithms provide acceptable results, autoencoders are more confident in their predictions in cases where alarm information cannot help so much with separating faulty from healthy turbines. Autoencoders, having more tunable parameters and allowing for more elaborated structures, are capable to better interpret non-linear data, such as that of a turbines generator. Additionally, the overall methodology was tested against a normality model, and the results clearly showed that the proposed solution ranks better for all the tracked statistics.

This research contributes to present a novel methodology that makes use of data analysis techniques for anomaly detection and consolidates the results, merging the anomaly predictions with information from the alarm system. The large size of the datasets and its diversity contribute to prove the approach as a general solution that can work well in real-life conditions and is not only applicable to a niche of turbines.

Different network architecture, including temporal information and denoising autoencoders, should be explored in future research to boost the accuracy of the system. Interpretability of results is a key aspect that requires further improvements to ensure acceptance of this methodology in the market.

Author Contributions: Conceptualization, M.B. and J.J.C.; methodology, M.B., J.J.C. and C.K.; software, M.B., J.J.C. and C.K.; validation, M.B., J.J.C., J.C. and C.K.; formal analysis, M.B., J.J.C. and J.C.; investigation, M.B.; resources, J.C.; data curation, M.B. and J.C.; writing—original draft preparation, M.B.; writing—review and editing, J.C.; visualization, M.B.; supervision, J.J.C. and J.C.; project administration, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors want to acknowledge the support of the other colleagues at SMARTIVE, Beatriz Jaramillo and Jose Sepulveda, for their support and technical advice in the design of this experiment and the analysis of the results. Additionally, the authors would like to acknowledge the work of the anonymous reviewers, who provided useful notes and critiques to improve the overall quality of the article.

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

References

1. Wind Energy in Europe in 2018—Trends and Statistics. 2019. Available online: https://windeurope.org/wp-382content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2017.pdf (accessed on 12 March 2020).
2. Faulstich, S.; Hahn, B.; Tavner, P.J. Wind turbine downtime and its importance for offshore deployment. *Wind. Energy* **2010**, *14*, 327–337. [CrossRef]

3. Making Transition Work. 2016. Available online: https://windeurope.org/wp-content/uploads/files/about-386wind/reports/WindEurope-Making-transition-work.pdf (accessed on 12 March 2020).
4. Schlechtingen, M.; Santos, I.F. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mech. Syst. Signal Process.* **2011**, *25*, 1849–1875. [CrossRef]
5. Gonzalez, E.; Stephen, B.; Infield, D.; Melero, J.J. On the use of high-frequency SCADA data for improved wind turbine performance monitoring. *J. Phys. Conf. Ser.* **2017**, *926*, 392. [CrossRef]
6. Ouyang, T.; Kusiak, A.; He, Y. Modeling wind-turbine power curve: A data partitioning and mining approach. *Renew. Energy* **2017**, *102*, 1–8. [CrossRef]
7. Shokrzadeh, S.; Member, S.; Jozani, M.J.; Bibeau, E. Wind Turbine Power Curve Modeling Using Advanced Parametric and Nonparametric Methods. *IEEE Trans. Sustain. Energy* **2014**, *5*, 1262–1269. [CrossRef]
8. Pandit, R.K.; Infield, D.; Kolios, A. Gaussian process power curve models incorporating wind turbine operational variables. *Energy Rep.* **2020**, *6*, 1658–1669. [CrossRef]
9. Teng, W.; Ding, X.; Zhang, X.; Liu, Y.; Ma, Z. Multi-fault detection and failure analysis of wind turbine gearbox using complex wavelet transform. *Renew. Energy* **2016**, *93*, 591–598. [CrossRef]
10. He, G.; Ding, K.; Li, W.; Jiao, X. A novel order tracking method for wind turbine planetary gearbox 405 vibration analysis based on discrete spectrum correction technique. *Renew. Energy* **2016**, *87*, 364–375. [CrossRef]
11. Mollasalehi, E.; Wood, D.; Sun, Q. Indicative Fault Diagnosis of Wind Turbine Generator Bearings 408 Using Tower Sound and Vibration. *Energies* **2017**, *10*, 1853. [CrossRef]
12. Dias, H.; de Azevedo, M.; Maurício, A.; Bouchonneau, N. A review of wind turbine bearing 410 condition monitoring: State of the art and challenges. *Renew. Sustain. Energy Rev.* **2016**, *56*, 368–379. [CrossRef]
13. Xu, H.; Feng, Y.; Chen, J.; Wang, Z.; Qiao, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; et al. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. In Proceedings of the WWW 2018: The 2018 Web Conference, Lyon, France, 23–27 April 2018; pp. 187–196.
14. Yousefi-azar, M.; Varadharajan, V.; Hamey, L.; Tupakula, U. Autoencoder-based Feature Learning 415 for Cyber Security Applications. *Int. Jt. Conf. Neural Netw.* **2017**, 3854–3861. [CrossRef]
15. Yan, W.; Yu, L. On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach. In Proceedings of the Annual Conference of the Prognostics and Health Management Society, San Diego, CA, USA, 19–24 October 2015.
16. Jiang, G.; He, H.; Xie, P.; Tang, Y. Stacked Multilevel-Denoising Autoencoders: A New Representation Learning Approach for Wind Turbine Gearbox Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2017**, *66*, 2391–2402. [CrossRef]
17. Jiang, G.; Xie, P.; He, H.; Yan, J. Wind Turbine Fault Detection Using Denoising Autoencoder with Temporal Information. *IEEE/ASME Trans. Mechatron.* **2017**, *23*, 89–100. [CrossRef]
18. Wang, Q. Intelligent wind turbine blade icing detection using supervisory control and data acquisition 427 data and ensemble deep learning. *Energy Sci. Eng.* **2019**, *7*, 2633–2645. [CrossRef]
19. Luis, M.L.; Rodríguez-l, M.A. Development of indicators for the detection of equipment malfunctions and degradation estimation based on digital signals (alarms and events) from operation SCADA. *Renew. Energy* **2016**, *99*, 224–236. [CrossRef]
20. Gonzalez, E.; Reder, M.; Melero, J.J. SCADA alarms processing for wind turbine component failure detection SCADA alarms processing for wind turbine component failure detection. *J. Phys. Conf. Ser.* **2016**, *753*, 072019. [CrossRef]
21. Sakurada, M.; Yairi, T. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, Gold Coast, Australia, 2 December 2014; pp. 4–11. [CrossRef]
22. An, J. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. *Spec. Lect. IE* **2015**, *2*, 1–18.
23. Marti-Puig, P.; Blanco-M, A.; Cárdenas, J.J.; Cusidó, J. Effects of the pre-processing algorithms in fault diagnosis of wind turbines. *Environ. Model. Softw.* **2018**, *110*, 119–128. [CrossRef]
24. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55. [CrossRef]

25. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
26. Li, K.; Huang, H.; Tian, S.; Xu, W. Improving one-class SVM for anomaly detection. In Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, China, 2–5 November 2003. [CrossRef]

4. Ensemble Learning Frameworks for SCADA Data

4.1 Ensemble Learning

The second topic tackled by this thesis is the application of ensemble learning to wind energy predictive maintenance. Two research papers —implementing ensemble models to predict main bearing failures— are presented. The data used is SCADA time series of multiple windfarms, and both machine learning and deep learning models were implemented.

Ensemble learning is a popular sub–domain of machine learning. Ensembles are made of the aggregation and combination of weak learners —simpler algorithms that not always fit perfectly the data— the objective is to reduce predictions variance while maintaining a low bias [92], [93]. This can be obtained using techniques based on simple voting schemes, or even more complex approaches based on meta–learners.

Meta–learning structures are algorithms that process the output of other algorithms to build a higher level model [94]. Data–science contests, such as the ones hosted on Kaggle [1], are contexts where ensemble learning shines and enjoys large popularity. In this competitions a very successful strategy is the combination of different algorithms to form an aggregated prediction. Under some conditions —namely low degree of correlation between the predictions— improved performance are often achieved. While this technique has been proven useful and successful in the resolution of a large array of problems in the data–science literature, it is not applied in this research.

The main limitation is the absence of a sufficiently reliable and high quality labelled dataset that makes challenging to train a robust meta–classifier. Instead, simpler solution based on ranking

[1] https://www.kaggle.com/

and combining predictions of base algorithms are preferred. This solution not only allows to deal with the scarce quality of labels, but it also leads to more intuitive results. The contribution of the different algorithms can be determined and the type of anomaly encountered in the turbine explained. This last characteristic is vital to gain the trust of turbines' maintainers and drive concrete corrective actions.

4.2 An Ensemble Learning Solution for Predictive Maintenance of Wind Turbines Main Bearing

This paper presents a predictive maintenance strategy for wind turbines' main bearing. The main bearing is a crucial component of a turbine's drive–train. The large dimension makes maintenance particularly troublesome; Cranes and specialized equipment is often needed for reparations and substitution. Early fault warnings are attractive to turbines' owners as they allow to optimize logistics, minimize downtime, and improve scheduling of corrective measures.

The chosen modelling framework is based on ensemble models which are able, not only of achieving good results, but also provide a very valid solutions for combining different data sources and heterogeneous inputs. In the predictive maintenance field using alarms and textual data in conjunction to time series can be challenging as the format and frequency of the data differ greatly. Time series are commonly stored using different flavors of numerical data types —most often comma separated values— alarm and work order logs are commonly provided as text data.

The previous chapter showed the advantages of mixing different data sources, now the focus is on the utilization of specialized algorithms that are able to capture particular traits of the same set of data. Ensemble learning allows to mix the output of the various specialized algorithms into a single, more reliable prediction.

The following indicators have been combined to assess main bearing conditions:

- *weekly mean indicator*
- *weekly normality indicator*
- *weekly anomaly indicator*

These are popular predictive strategies encountered in the literature. Algorithms have been adapted to capture patterns typical to data prior the occurrence of main bearing failures, such as increasing trends in the bearing temperature.

The results have been obtained monitoring data of two windfarms, and a total of 84 turbines for a period of approximately two years. Performance metrics, i.e. accuracy, precision and F1 score have clearly benefited from the combination of indicators. Predictions have been served to windfarm owners; The practical and pragmatic nature of the indicators were appreciated features which helped removing the aura of mystery that predictive algorithms often have.

Contributions

The main novelties and contributions of this paper include:

- **Design** of an unsupervised process which makes limited assumptions on data. SCADA data can be used for predictions and labels are not required, thus greatly simplifying the preprocessing pipeline.
- The algorithm can be **easily scaled**, single or multiple windfarms can be monitored without the need to redesign the entire predictive framework.
- **Demonstrate** that ensemble learning is an effective strategy to combine predictions of various interpretable indicators, capturing diverse characteristics of the signal.

sensors MDPI

Article

An Ensemble Learning Solution for Predictive Maintenance of Wind Turbines Main Bearing

Mattia Beretta [1,2], Anatole Julian [2], Jose Sepulveda [2], Jordi Cusidó [2,3,*] and Olga Porro [2,4]

[1] Unitat Transversal de Gestió de l'Àmbit de Camins UTGAC, Universitat Politécnica de Catalunya, 08034 Barcelona, Spain; mattia.beretta@smartive.eu

[2] SMARTIVE S.L., 08204 Sabadell, Spain; anatole.julian@smartive.eu (A.J.); jose.sepulveda@smartive.eu (J.S.); olga.porro@smartive.eu (O.P.)

[3] Enginyeria de Projectes i de la Construcció EPC, Universitat Politécnica de Catalunya, 08028 Barcelona, Spain

[4] Facultat de Matemàtiques i Estadística, Universitat Politécnica de Catalunya, 08028 Barcelona, Spain

* Correspondence: jordi.cusido@smartive.eu

Abstract: A novel and innovative solution addressing wind turbines' main bearing failure predictions using SCADA data is presented. This methodology enables to cut setup times and has more flexible requirements when compared to the current predictive algorithms. The proposed solution is entirely unsupervised as it does not require the labeling of data through work orders logs. Results of interpretable algorithms, which are tailored to capture specific aspects of main bearing failures, are merged into a combined health status indicator making use of Ensemble Learning principles. Based on multiple specialized indicators, the interpretability of the results is greater compared to black-box solutions that try to address the problem with a single complex algorithm. The proposed methodology has been tested on a dataset covering more than two year of operations from two onshore wind farms, counting a total of 84 turbines. All four main bearing failures are anticipated at least one month of time in advance. Combining individual indicators into a composed one proved effective with regard to all the tracked metrics. Accuracy of 95.1%, precision of 24.5% and F1 score of 38.5% are obtained averaging the values across the two windfarms. The encouraging results, the unsupervised nature and the flexibility and scalability of the proposed solution are appealing, making it particularly attractive for any online monitoring system used on single wind farms as well as entire wind turbine fleets.

Keywords: main bearing; wind turbine; failures; predictive maintenance; ensemble learning; unsupervised; interpretable; scalable; SCADA

check for updates

Citation: Beretta, M.; Julian, A.; Sepulveda, J.; Cusidó, J.; Porro, O. An Ensemble Learning Solution for Predictive Maintenance of Wind Turbines Main Bearing. *Sensors* **2021**, *21*, 1512. https://doi.org/10.3390/s21041512

Academic Editor: Steven Chatterton

Received: 3 February 2021
Accepted: 18 February 2021
Published: 22 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The future is bright for wind energy. New turbines are being installed, technologies are improving and costs are decreasing. IRENA estimates a tumultuous growth for the industry, expecting a global installed capacity of 1000 GW by 2050, and new installations rate of 200 GW/yr, including replacement of old turbines [1]. By the end of 2019 Europe alone boasted 205 GW of installed wind power capacity [2].

A number of challenges have to be faced in order to reach such ambitious goals, reducing costs of operation and maintenance (O&M) is paramount. In large wind farms O&M costs can account up to 30% of the total cost of energy, the influence of physical maintenance is estimated around 20% of the levelized cost of electricity (LCOE) [3]. Turbines are often situated in remote locations, the components are bulky and difficult to transport, logistics costs are significant. The growth of offshore installation, which accounted for 22 GW of power capacity in Europe in 2019 [2], exacerbates the problem as logistics becomes even more challenging.

Of all the components in a turbine: the main bearing, which provides support to the ma axis connecting blades and gearbox, is one of the most problematic in terms of maintenance and logistics. Failure rates reaching 30% for single main-bearing and of 15% for double main-bearing turbines, over 20 year lifetime are reported by Hart et al. [4]. Replacing a main bearing is no trivial task, unlike other systems that can be repaired in on-tower interventions, a crane is needed and the faulty turbine has to be put out of production for a long period of time.

Most turbines are equipped with a Supervisory Control and Data Acquisition (SCADA) system. This is a network of sensors monitoring various physical quantities: such as temperature, speed and pressure of the principal components of a turbine. International Standards, such as IEC 61400-25 simplify the representation of turbines and guarantee the uniformity of information exchange and control design [5]. While initially designed for control purposes only, SCADA data have also predictive capabilities and it has been used widely in the literature [6,7].

This articles presents a solution built on SCADA data to address main bearing failures, predicting the occurrence of future faults, and thus, helping wind farm operators to improve maintenance and reduce costs related to unexpected failures. Predictions from a set of understandable indicators, designed to capture different characteristics of the signal, are combined into a composed health status indicator. Data from two onshore wind farms, for a total of 84 monitored turbines, is used to evaluate the performances of this solution.

The main contributions of this research can be summarized in three key-points:

1. Present an unsupervised system, requiring minimum setup and limited prerequisites, capable to monitor entire wind farms.
2. Provide interpretable and understandable predictions, in contrast to black-box solutions.
3. Implement an Ensemble Learning strategy that produces reliable predictions from a set of understandable indicators, improving their individual performances.

1.1. Main-Bearing Failure Discussion

The rolling elements of wind turbines' main-bearing are subjected to severe working conditions, far different from the typical stress that are known in other industrial applications such as power plants. Windspeed, turbulence index and in general variations of the wind field conditions have a significant effect on main bearing deterioration [4].

The principal damage and wear mechanisms are reported by Hart et al. [8], defects in the assembly, design and manufacturing of main bearings lead to premature wear of the main bearing. Phenomenons such as micro-pitting, spalling, smearing etc. can be observed [8]. Progressive wear of material leads to sub-optimal operating conditions, higher localized loads where defects arise and in general overheating of the main bearing.

An incipient main-bearing failure is expected to be preceded by anomalous vibrations and increases in temperature of the component. In this study, vibration measurements are not available, thus the attention is given to anomalous patterns in temperature readings. Moreover, temperature signals are easy to interpret and they are part of the typical recordings of a SCADA system, unlike vibration signals that rarely are available. The use of temperature data thus make this solution applicable for a wider range of wind-farms.

Different authors have successfully studied temperature behaviors to predict failures in various turbines' components. Guo et al. devised a monitoring strategy for turbines' generators based on tracking of the generator temperature via change detection of a memory matrix of the component behavior [9]. Qiu et al. presented a thermophysics approach to assess drive train conditions from which various diagnostic rules are defined [10]. Tonks and Wang showed experimentally that monitoring temperature can reveal misalignments and problems of shaft couplings, as these defects increase friction therefore temperature of the component [11]. Cambron et al. developed a method to monitor main bearing condition comparing the measured and expected temperture of the component, predictions were obtained using a physical model of the bearing [12]. Sun et al. describe an anomaly identification method using mainly temperature readings and other standard SCADA signals to monitor the behavior of the major components [13].

One of the main bearing failure event is presented. Figure 1 shows the temperature profile of the faulty turbine and the average of the wind farm. The damaged main bearing is evidently warmer than the average. In Figure 2, other evidences of the failure are visible, the relation between main bearing temperature and wind speed is steeper for the defective turbine. Moreover, the density distribution of the faulty main bearing is shifted to higher values than the wind-farm average.

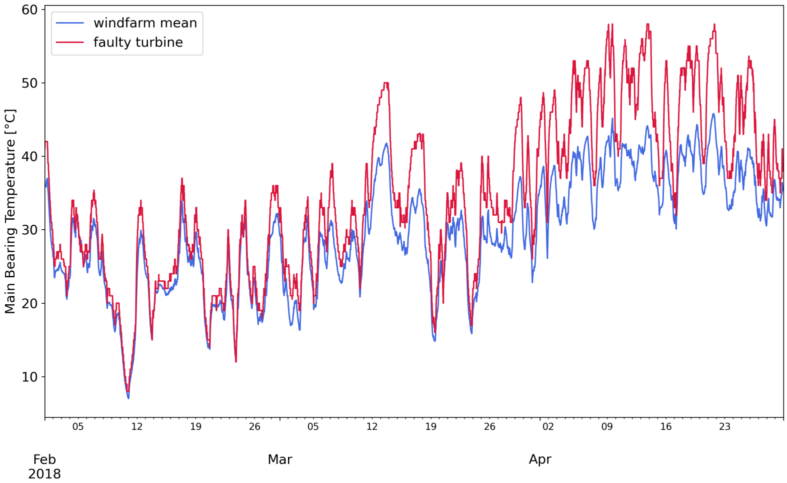


Figure 1. Timeseries profiles of the main bearing temperature of a faulty turbine and the average of the wind-farm.

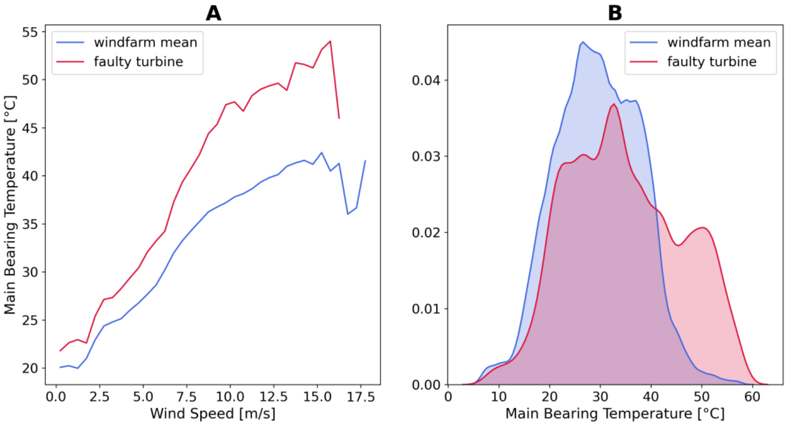


Figure 2. (**A**) Relation between main bearing temperature and wind speed. (**B**) Probability density plot of the main bearing temperature of a faulty turbine and the average of the wind-farm.

The paper is organized as follows. Section 2 is a review of previous works available in the literature. Section 3 provides an explanation of the data used and the applied pre-processing techniques. Section 4 details how the solution is built, showing the base components and how they are combined into a single health status indicator. In Section 5, results are presented and analyzed, followed by Section 6, where a discussion is provided. Finally, Section 7 contains the final remarks and recommendations for future work directions.

2. Previous Works

Various solutions are available to assess the status of wind turbine components and predict failures. Methods can be classified by the type of data utilized. Vibrations, currents and acoustics measurements are particularly effective to diagnose drive-train failures, as documented in [14–18]. These solutions require the installation of additional sensors or in-situ measurements campaigns to collect the data. On the contrary, SCADA system is available as standard equipment for most turbines and its recordings registered in the databases of wind farm owners, such that operators who did not think in advance of data-based predictive maintenance strategy can implement one, using SCADA data, at minimal additional costs.

SCADA predictive maintenance algorithms can be sorted in multiple categories as proposed by Tautz and Watson [6]. In this paper, the following are analyzed:

1. Signal Trending;
2. Normality models;
3. Anomaly detection and Clustering methods.

These three methods are reviewed in the following Sections 2.1–2.3. Then, a review of Ensemble Learning is provided in Section 2.4 since this is an essential component of the methodology. Relevant applications in predictive modeling and data analysis are discussed. Gradient Boosting and Isolation Forest are also presented in Sections 2.5 and 2.6, respectively, as they are used in our solution.

2.1. Signal Trending

The signal trending approach is based on the study of changes and trends in a long period of time. The underlying hypothesis of this approach is that failures have a sort of signature that can be detected observing variables such as temperatures.

Astolfi et al. proposed a simple, but effective methodology to monitor turbine components. The relation between binned active power and key sensor's readings such as rotor and generator bearing temperature are tracked within the wind-farm and through time obtaining useful visualization of the state of the turbines and an effective failure detection tool [19]. Cambron et al. proposed a control chart monitoring algorithm based on the comparison of turbines against wind-farm average to detect problems in the generator [20]. Yang et al. presented a technique to track incipient failures through the analysis of the relation between some key variables and contextual parameters such as the wind speed, as shown in the two case-studies the progression of failures is gradual through time and trends towards anomalous conditions can be observed [21]. Feng et al. devised a failure detection strategy for gearboxes based on the thermodynamics and physical behavior of this component, a relation between the loss of efficiency and increase in temperature is derived and utilized to analyze a known failure [22]. Li and Yu formulated a method based on the difference of the median of each turbine with the rest of the wind-farm and used it to build a condition vector. The authors use monitoring charts to generate alarms and discuss several strategies to deal with autocorrelation of operation data [23].

Main advantages of these methods are: ease of implementation, straightforward interpretation of the results and limited data requirements. Being based on simple statistics they can be replicated with minimal knowledge of advanced algorithms and data-analysis techniques. Moreover, the underlying hypotheses of these methods are rooted in the thermodynamics and physical principles governing operations of the components. Wind-farm maintainers often track the same deviations and trends that are automatized by these algorithms, thus results will sound familiar and understandable.

That being said, many of these methods are univariate and are not capable of capturing the interactions between multiple variables. Being wind turbines complex systems, based on the interconnection of mechanical, electrical and electronic components this limitation can be significant. Moreover, incorporating the influence of external variables, such as wind speed and external temperature is not trivial for these methods.

2.2. Normality Models

Normal Behavior Modeling (NBM) is a class of predictive algorithms attempting to infer the relation between a set of inputs and a target variable under normal operation of a turbine component. Deviations between predictions and measurements of the target sensor are used to detect failures.

Schlechtingen and Santos compared simple regression models to more sophisticated implementations based on neural networks; details on the training and utilization of normality models are also provided [24]. Puig et al. presented a normality model for turbine generator and gearbox based on Extreme Learning Machines that can be deployed in the cloud, allowing real-time operations [25]. Zhang and Wang proposed an artificial neural network solution for fault detection in wind turbines main bearings, using SCADA data and able to anticipate failures, allowing to schedule maintenance avoiding unexpected breakdowns [26]. A self-evolving maintenance scheduler, based on artificial neural network tracking gearbox bearings conditions is discussed by Bangalore and Tjerrnberg [27]. Normality models are a well established solution in wind turbines' predictive maintenance field.

The multivariate nature of this approach is suited to capture complex relations between turbines' sensors, advanced algorithms and neural network architectures can be used to detect non-linearities in the data and model turbine behavior.

Two main criticisms can be addressed to normality models. First, the interpretability of the predictions is scarce as often sophisticated algorithms are used and the influence of input parameters on the output prediction is not trivial, the behavior is that of a 'black-box'. Second, the selection of the training set to feed to the algorithm is crucial. This task is time-consuming, the sample of data should include all possible operating and external conditions, thus training set shorter than one year are not particularly reliable. On top of that, normal operating conditions only should be selected, this involves a thorough analysis of the turbines logs and eliminations of alarms and unusual operating instances.

2.3. Anomaly Detection and Clustering Methods

Anomalies in SCADA data can be detected modifying NBMs. Instead of predicting the value of a target variable using regressive models, the physical model underlying input variables can be learned and the difference between the original and reconstructed signal tracked. Autoenconders (AE), Restricted Boltzmann Machines (RBM) and Generative Adversarial Networks (GAN) are suited for this task [28–30]. Signal reconstruction algorithms are capable of capturing non-linearities and produce refined models of the data. On the other hand, as for NBMs, a training set composed of normal operation data is needed. Moreover, complex structures such as AE and GANs often require large volumes of data.

Clustering offers an alternative approach, data is analyzed in search of meaningful groups that can capture interesting relationships within the input variables. Blanco et al. presented a methodology based on Self-Organizing Maps (SOM) and clustering to assess wind turbines' health status [31]. Du et al. also proposed a SOM based solution to identify system level anomalies [32]. These methods are able to produce insightful representations of the data, that can help the analyst to discover unexpected, but interesting relationships. The purely unsupervised nature though, leads to significant problems in the integration of these algorithms in automatic predictive pipelines. Rules, thresholds and other solutions are needed to make these solutions valuable in an online system.

A large selection of Machine Learning algorithms can also be used for anomaly detection. McKinnon et al. have studied the performances in condition monitoring of a gearbox of three popular algorithms: Isolation Forest (IF), One Class Support Vector Machine (OCSVM) and Elliptical Envelope (EE) and found that depending on the conditions OCSVM and IF reach best results [33]. Purarjomandlangrudi et al. used Support Vector Machine (SVM) to process previously extracted features of the data for early detection of anomalies [34]. Isolation Forest is a particularly interesting approach as it does not require a normal operation dataset to characterize data, anomalies are determined analyzing the density of data in the different regions of the feature space [35]. On top of that, these methods can deal with multivariate distributions and normally require less data and training time with respect to more complex Deep Learning solutions.

2.4. Ensemble Learning

Predictions of base learners, sufficiently independent from each other, can be combined into a meta-predictor which often achieves better performances than the individual predictors. This approach is typically referred to as Ensemble Learning, some of its declinations are: boosting, bagging, model averaging and stacking.

This learning paradigm is particularly popular in data-science competitions, a famous example is the algorithm that won the "Netflix Challenge" [36]. An example of Ensemble Learning in an industry application is presented by Wu et al. that used ensembling to deal with imbalanced datasets [37]. A meta-learner trained on a subset of base predictors has been used to improve wind power production in [38,39]. Liu et al. proposed a solution to detect wind turbine blades icing combining features extracted by Deep-Autoencoders into an ensemble model where decision is taken by majority vote [40]. Ensembles can be used to merge information from different data sources, as Turnbull et al. demonstrated using a OCSVM to combine NBMs of a temperature SCADA and vibration data for gearbox and generator bearings of wind turbines [41].

Most of the aforementioned literature make use of a meta-algorithm trained on the predictions of base learners. To do so, a subset of the data have to be withhold to train the higher level algorithm and adjust its parameters. Work orders are used to label healthy and faulty operating conditions of turbines. In this research, an alternative approach is taken, instead of training a high-order classifier, the predictions of the individual unsupervised algorithms are combined into a single health status indicator, to avoid the necessity of labeling data.

2.5. Gradient Boosting

First introduced by Friedman, gradient boosting machine is a popular Ensemble algorithm applied both in classification and regression problems [42]. This technique makes use of base-learners, typically decision trees, to learn the relation between input and output data.

The algorithm is iterative as new base learners are routinely trained on a dataset. The name gradient boosting encapsulates the key idea of this technique: accelerating the convergence towards the optimum set of parameters that minimizes the adopted loss function.

Concretely, at each new iteration residuals between prediction and real values are calculated and larger weights are assigned to the instances where the error is greater such that more efforts will be made to fit the model to them. The process is repeated until a stopping criteria, such as the maximum number of iterations or the minimum error, is reached. The algorithm from the original paper [42] is reported below.

ALGORITHM: Gradient Boosting

Given input data $(x, y)_{i=1}^{N}$, a differentiable loss function $L(y, \rho)$, a base learner $h(x, \mathbf{a})$, a function $F(\mathbf{x})$ to estimate and a maximum number of iterations M.

These are the steps to follow:

```
1. $F_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, \rho)$
2. For $m = 1$ to $M$ do:
3. $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(\mathbf{x})}, i = 1, N$
4. $\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$
5. $\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$
6. $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$
7. endFor
   end Algorithm
```

The algorithm works with a large selection of loss functions and guarantees short training and predicting times. Variations such as XGBoost and LightGBM exist to address some shortcomings of the original algorithm, granting parallel execution and more tuneable parameters.

2.6. Isolation Forest

This algorithm was introduced in 2008 by Liu et al. [35]. The founding principle of this method is that anomalies are usually a minority within the data and can be easily divided from the rest of the dataset. With this in mind, multiple fully developed randomized trees are fully trained, meaning that each of their terminal leaf is to be composed of one point only.Trees splits are made setting a random threshold, instead of the optimal one.

Being an Ensemble method this procedure is repeated multiple times, training an entire forest of decision trees. The average path length, meaning the number of splits necessary to isolate a given point, is used to define an anomaly score defined in Equation (1):

$$s(x, \psi) = 2^{\frac{-E(h(x))}{c(\psi)}} \tag{1}$$

where $E(h(x))$ is the average value of the path length for a given point, $c(\psi)$ is the average path length of unsuccessful search in Binary Search Trees and ψ number of instances. Values of s approaching 1 are related to anomalies, scores lower than 0.5 are associated with normal observations and finally, if the entire dataset has scores close to 0.5 no evident anomalies are present.

3. Data

SCADA data (10 min time resolution) of two onshore wind farms are used. More than two years of operation are analyzed for a total of 84 turbines. The first wind farm, located in North America, is made of 66, 1.5 MW rated power turbines; the second one, situated in Poland, has 18, 2 MW turbines. SCADA data comes in comma-separated values (csv) format files. The dataset and pre-processing steps are discussed in the following subsections.

3.1. SCADA Dataset

The original SCADA dataset is composed of hundreds of columns, since turbines are typically equipped with a multitude of sensors monitoring various components. These sensors record the state of the system at a high frequency. Then, they are downsampled to lower resolution, most commonly 10 min. Raw signal is summarized by taking its mean, standard deviation, minimum and maximum value during the aggregation period. An example of the SCADA dataset is presented in Table 1. In this research, only the main bearing temperature sensor, active power output, environment temperature, wind speed and rotor speed are used, reducing significantly the dimensionality of the dataset. The choice of these variables is dictated by the necessity to characterize the main-bearing working conditions and the context in which it is operating. The relevance of the variable selection has been certified by experts of the wind turbine maintenance field.

Table 1. Sample of SCADA data.

Turbine	Timestamp	Main-Bearing Temp. C°	Active Power W	External Temp C°	Wind Speed m/s	Rotor Speed rpm
WT01	02/01/18 10.00 am	32	1529	−6	14	17
WT01	02/01/18 10.10 am	32	1532	−6	13	17
WT01	02/01/18 10.20 am	32	1532	−6	13	17

3.2. Data Processing

Real-life data is typically affected by missing records or outliers, caused by miss-communications or defects of the sensors. A preliminary filter of absurd readings is necessary to reduce the chances of generating false alarms. In the Literature various data filtering approaches have been proposed, most of them are based on the application of statistical filters [43]. In this research a manual threshold values based on technical knowledge of turbines behaviors are used to filter data, as the number of variables to analyze is limited. Values trespassing the imposed thresholds have been removed from the dataset, no imputation nor interpolation are used to fill the gaps.

4. Methodology

The scheme of the proposed solution is illustrated in Figure 3. The three indicators used to analyze the data are the following: Mean average temperature of the main bearing; Normality model; and Anomaly detection algorithm.

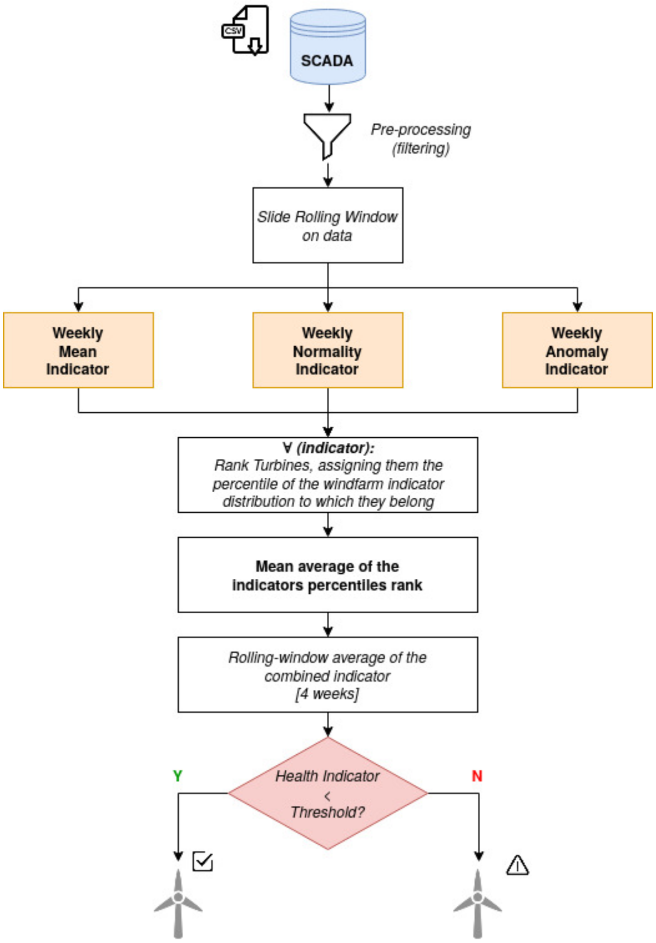

CSV
SCADA
Pre-processing (filtering)
Slide Rolling Window on data
Weekly Mean Indicator
Weekly Normality Indicator
Weekly Anomaly Indicator
∀ (indicator): Rank Turbines, assigning them the percentile of the windfarm indicator distribution to which they belong
Mean average of the indicators percentiles rank
Rolling-window average of the combined indicator [4 weeks]
Health Indicator < Threshold?
Y
N

Figure 3. Diagram of the predictive maintenance solution.

Each indicator is calculated from raw data at 10 min resolution, using the rest of the wind farm as meter of comparison, a similar approach is used in [19,23,44]. Turbines belonging to the same wind farm are typically from the same manufacturer and technology. Moreover, with regard to external conditions, measurements registered at each turbine such as wind speed and external temperature behave similarly for a given period of time. Results are aggregated on a weekly basis to account for timely variation of conditions between turbines that could skew results excessively. The decision of the weekly aggregation time-frame is dictated by a compromise between ensuring continuous and precise monitoring of turbines and avoiding to flood maintainers with updates on the wind-farm status. The final assessment of the main bearing status is given by the comparison between the averaged value of the combined indicator over a 4 week period and a decision threshold.

A sliding window, as shown in Figure 4 is used to scan the data. On the left side, the normality models rolling scheme, train and test sets are illustrated. On the right side, the rolling window used for the the other two indicators, whose output is calculated directly on the analyzed data, without the need of a training phase, is shown.

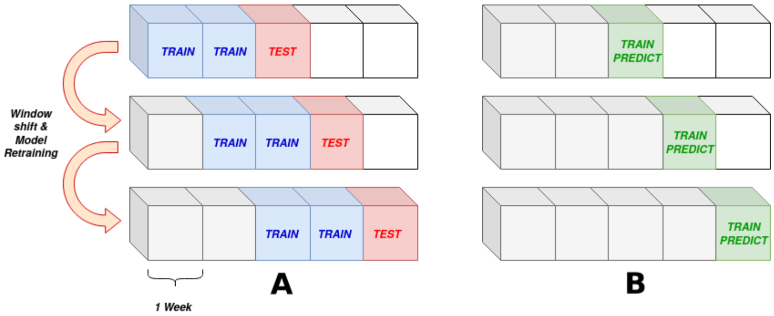


Figure 4. (**A**) The rolling window train/test scheme used for normality models. (**B**) The rolling window train/test scheme used for mean and anomaly indicators.

4.1. Mean Average Indicator

The first indicator tracks the weekly mean average temperature of turbines' main bearing. This indicator is used to determine whether some turbines are operating at consistently higher temperatures with respect to the wind farm. As presented in Section 1.1, higher temperatures of the main bearing are a common pattern in faulty turbines. An example of the temperature distribution of main bearings is presented in Figure 5. Variation between the turbines is evident.

This indicator is straightforward and easy to interpret, but being the measure of a univariate series, it cannot account for crossed relations between variables such as different operating conditions of the turbines. Higher temperatures may be caused simply by higher production conditions.

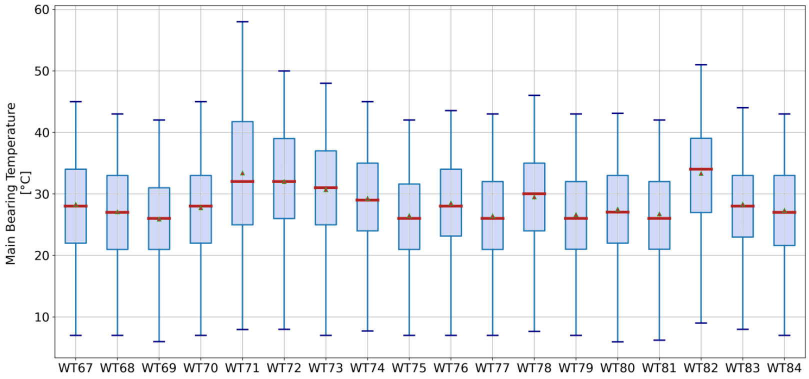


Figure 5. Boxplot of the main bearing temperature. The median is represented by the red line and the mean corresponds to the triangle.

4.2. Normality Model

Normality models are used to infer the relation between some inputs and a target variable, that can characterize the system under analysis. Normal operating data is needed to train the algorithm and infer the expected behavior of the system. The trained model can be used to predict values that are compared to the measurements of the target variable. Large deviations between predicted and observed values are to be considered suspicious, as they represent deviations from normal behavior.

The pre-selection of normal data is a time-intensive task as it requires the analysis of the work order logs to remove faulty data and abnormal conditions. Automating this task is not trivial and retrain is needed after repairs and modifications of the component. This research presents an adaptation of normality models that allows to skip the labeling step, reducing greatly time overheads in the training phase of the model.

A rolling window, as the one shown in Figure 4 is slid over data, its size being 8 weeks for the training set and 1 week for the test set. The window is then shifted by intervals of one week for next predictions. Instead of mapping the normal behavior of the turbine, the recent relation between the input and target variables is inferred during the training phase.

Deviations in this case, help to detect drifts in the target variable distribution as this is a pattern observed in main bearing failures. Obviously, difference between prediction and observed records can be the consequence of external conditions (high winds, heat waves, etc.) novel to the train set, in this case though a systematic error is expected in all turbines and alarms are unlikely to be raised, as all turbines will have large deviation.

The inputs used for this algorithm are:

- Active power [W];
- Wind speed [m/s];
- Rotor speed [rpm];
- External temperature [°C].

The main bearing temperature [°C] is used as output.

The sklearn implementation of gradient-boosting regressor for Python programming language is used [45,46]. The number of trees is set to 100 and their depth limited to 2, all other parameters are left to their default values. These parameters are found running cross-validation trials on a subset of the data. Deviation between a predicted and an observed value is measured calculating the root-mean squared error (RMSE), defined by Equation (2), where $\hat{y}_i$ and y_i are the predicted and the measured value, respectively, and N is the number of instances analyzed:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{2}$$

An example of the predictions for a given week and the RMSE by turbine is presented in Figure 6. Error is not uniformly distributed for the different operating conditions, what is important though is the comparison within the wind farm. Turbines that deviate more are isolated from the rest.

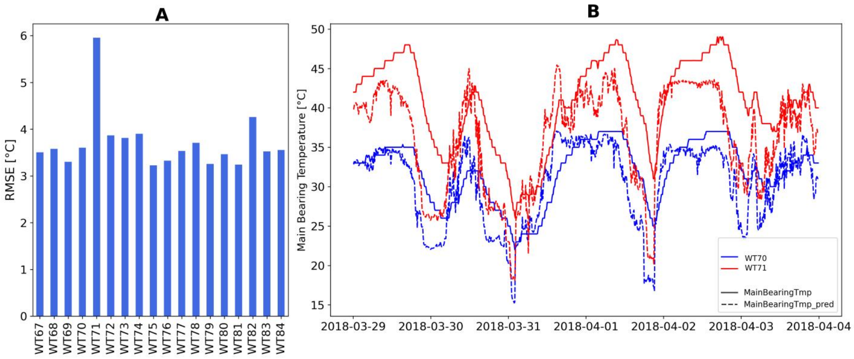


Figure 6. (**A**) Normality indicator, RMSE by turbine. (**B**) Timeseries comparison of predicted versus measured value.

4.3. Anomaly Detection

Isolation forest algorithm is used to detect anomalies in the windfarm data. Unlike other indicators that model turbines independently, the whole windfarm is analyzed at once with the objective to determine turbines that are behaving differently from the rest.

The feature space is composed by:

- Rotor speed [rpm];
- External temperature [°C];
- Main bearing temperature [°C].

Sklearn implementation of isolation forest is used [47], the percentage of anomalies is set to 10% of the data. This value is chosen after a series of tests on the sample of data. Choosing a higher percentage of anomalies will result in a larger number of normal points being considered as anomalies. A low value, instead, would lead to the isolation of very anomalous working conditions, missing other that can be relevant. A different dataset might require another value for this parameter, thus test of various values and examination of the indicator results are warmly recommended.

As for the other indicators, anomalies are calculated on a rolling-fashion, following the train-predict shown in Figure 4. Once anomalies are found, the percentage of anomalous records with respect to the total number of records for each turbine is calculated, see Equation (3), where AS_i is the anomaly score of turbine i, $\hat{x}_i$ is the number of anomalous points found for this turbine and x_i is the total number of points of the turbine.

$$AS_i = \frac{\hat{x}_i}{x_i} \tag{3}$$

This value is the anomaly detection indicator shown in Figure 7. Turbines having high percentage of anomalies are behaving differently with respect to the wind farm, thus should be more reasonably suspected to have some sort of problem. The right side of Figure 7 illustrates how isolation forest tends to separate data lying in peripheral regions of the feature space, where density of points is typically lower. On the left side, the percentage of anomalous points in each turbines for a given week is shown.

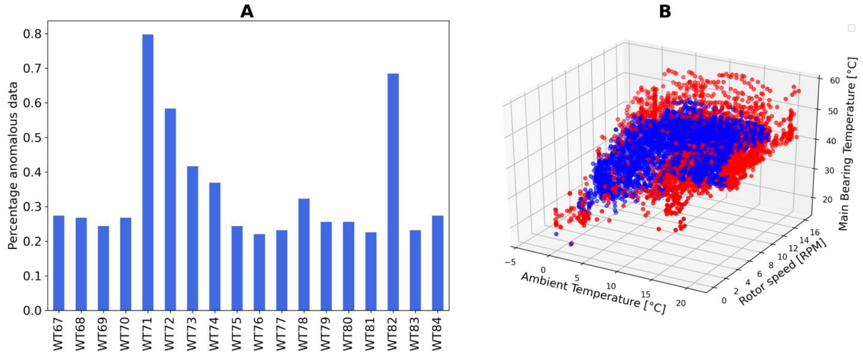


Figure 7. (**A**) Anomaly Indicator plots: percentage of anomalous versus total number of points. (**B**) 3D plot showing normal (blue) versus anomalous (red) points.

4.4. Indicators Merge Processing

The results of the individual algorithms are merged, obtaining a composed score of the turbine status. For each indicator is created a weekly ranking, assigning the percentile of the wind-farm distribution in which each turbine falls.

The three algorithms are designed to assign higher values to turbines, that according to their definition are to be considered faulty. The composition of the three values is calculated using a rolling average, with a sliding window of size 4 weeks as shown in Figure 8, using Equation (4). Where x_{ij} is the value of indicator j for a given turbine in week i.

$$H_{ind} = \frac{1}{N_{week} N_{ind}} \sum_{j=1}^{N_{ind}} \sum_{i=1}^{N_{week}} x_{ij} \tag{4}$$

Once the composed score is found, a decision threshold that decides if maintenance is defined. Setting the threshold is a trade-off between anticipating failures and having to do more maintenance intervention. A cost-benefit analysis is recommended to set this value to the value that maximizes economic savings, due to lack of information of the specific costs it has not been possible to optimize in such a way this parameter. A sensitivity analysis of the results is proposed instead.

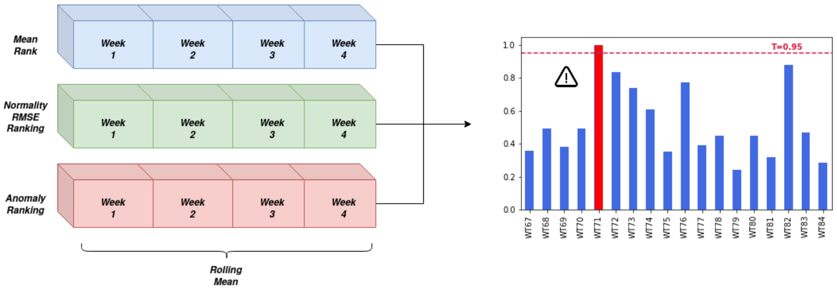


Figure 8. Composed indicator calculation scheme and decision threshold setting.

5. Results

Predictions for roughly two years of data are made and evaluated using the work orders logs. Windfarm operators commonly keep track of the checks and interventions required by the turbines. Unlike SCADA datasets, work orders logs do not follow standard formats. Records are typically organized as free-text. The time of the intervention, as well as the affected turbine and information regarding the actions taken are reported. Often, work order logs are used to filter data, removing abnormal operating conditions and assigning a healthy/faulty status to turbines. This research avoided this step, as the absence of a common standard makes difficult to automatize the labeling process; unsupervised algorithms have been favored instead. Work orders have been used only to assess the veracity of the predictions. The work order logs of the failures occurred during the period of analysis is presented in Table 2.

Table 2. Main Bearing failures work order logs.

Wind Farm	Location	Failure Date	Turbine	Comment
1	US	7 October 2017	WT31	Main Bearing Replacement
1	US	24 March 2018	WT62	Main Bearing Replacement
2	Poland	11 June 2018	WT71	Main Bearing Exchange
2	Poland	15 July 2019	WT72	Main Bearing Exchange

A limit to the anticipation period is defined, as an alarm is useful in practical terms only if it anticipates failures by a margin of time that allows wind farm operators to organize the replacement of the main bearing, optimizing the logistics and minimizing energy losses due to unexpected stops of the turbine. Weekly predictions are grouped in blocks of 4 months, if one alarm occurred during this period the turbine is reported for a maintenance check.

Performance of the proposed methodology is assessed by a confusion matrix. Predictions are sorted in the following categories:

- True Positive (*TP*);
- False Positive (*FP*);

- False Negative (*FN*);
- True Negative (*TN*).

A *TP* is assigned whenever an alarm is raised and the work order log reports a problem with the main bearing, if no problem is detected a *FP* is marked instead. In case a failure occurs and no alarm is raised, a *FN* is assigned. Finally, when no failure occurs and no prediction is given a *TN* is assigned.

5.1. KPIs Definition

A selection of performance indicators is used to track results, namely: accuracy, precision and F1 score. Their definition is defined using Equations (5), (6) and (7), respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FN + FP)} \tag{7}$$

5.2. Decision Threshold Sensitivity Analysis

As mentioned in the methodology, the decision threshold is an important parameter. It has a great influence on the results. A sensitivity analysis is proposed, in which the dependence of KPIs with respect to the decision threshold value is studied. The results of this analysis in the two wind-farms are shown in Figure 9.

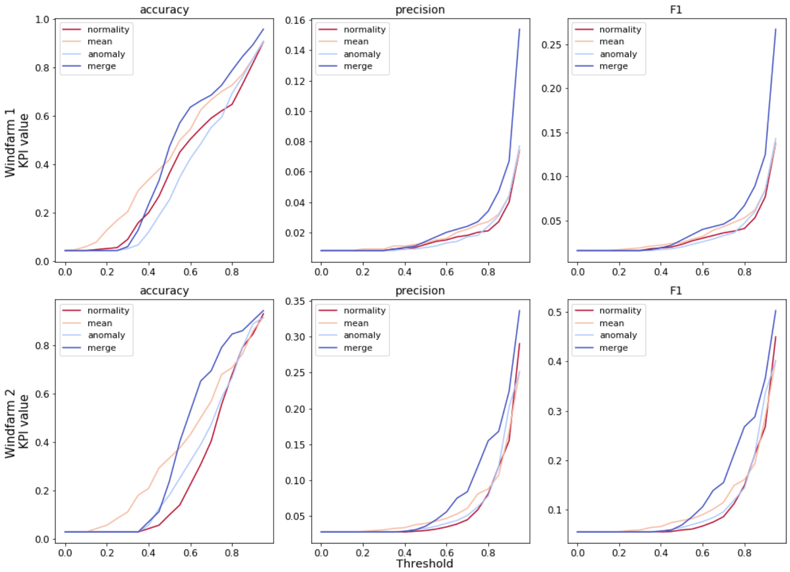


Figure 9. Relation between KPIs and decision threshold value by wind-farm and indicator.

Firstly, it should be observed that merging the information of the three indicators generally leads to improved performance, regardless of the decision threshold. Except for low values of the threshold, that have no practical relevance, since they would lead to an excessive number of reviews of the turbines.

Secondly, the algorithms are able to separate faulty turbines from healthy ones such that high decision threshold can be set. A high decision threshold means that only the most critical turbines will need checks and most of these reviews lead to the discovery of relevant problems, rather than false alarms.

That being said, a rigorous evaluation of the benefits and costs of choosing a certain value for the decision threshold is recommended to wind farm operators interested in this predictive algorithm. The cost of false alarms and unnecessary checks should be compared to the savings of early fault detection of a main bearing, and an economic optimum searched.

5.3. Comparison of Individual and Composed Indicator

The combination of the predictions of multiple algorithms leads to a better overall performance and this is one of the main claim of this research. This observation has been utilized in multiple fields of research, but not frequently by the wind energy predictive maintenance community. Having observed Figure 9, the decision threshold is assigned a value of 0.95 and a comparison of the available indicators and their composition is presented in Figure 10 and Table 3.

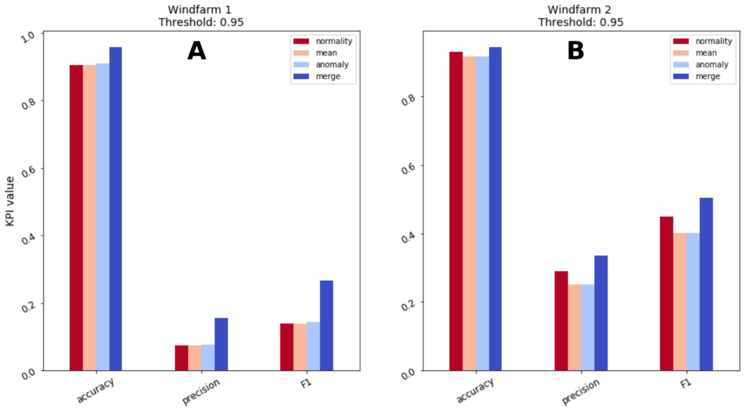


Figure 10. (**A**) Performance comparison of individual and composed indicators for Windfarm 1 and (**B**) Windfarm 2.

Table 3. Comparison of the results of individual and combined indicator for a threshold value of 0.95.

Windturbine	Indicator	TP	FP	FN	TN	Accuracy	Precision	F1
WF1	normality	48	600	0	5688	0.905	0.074	0.138
	mean	48	600	0	5688	0.905	0.074	0.138
	anomaly	48	576	0	5712	0.909	0.077	0.143
	merge	48	264	0	6024	0.958	0.154	0.267
WF2	normality	49	120	0	1577	0.931	0.29	0.45
	mean	49	146	0	1551	0.916	0.251	0.402
	anomaly	49	146	0	1551	0.916	0.251	0.402
	merge	49	97	0	1600	0.944	0.336	0.503

Combining predictions of individual indicators into a composed predictor is beneficial according to all the tracked metrics. Precision and F1 score benefit greatly from the combination of the indicators. For wind farm 1, precision and F1 scores double with respect to each single indicator as an effect of decreased number of FP, combining various sources allows to discard behaviors that are unusual, but not so critical to deserve maintenance check. Wind farm 2 also manifests an increase of precision and F1, but not as large as wind farm 1, overall results are better though as a precision of 33.6% and F1 score of 50.3% are reached. Accuracy is the metric that less benefits from the merging process as the starting values are already high, but an increase of 3–5 percentage points is recorded.

The information fusion process increase complexity of the predictive algorithm, but grants improved performance. Moreover, the design of simpler and specialized algorithms that focus on the detection of specific patterns in the data helps interpretability of the predictions. Base algorithms are implemented with the objective of capturing a specific trend in the data, rather than searching generic relationships within the variables. Once an alarm is raised the analyst can assess which indicators have greater influence in the alarm and verify whether the prediction is reasonable and eventually schedule a check of the turbine.

Information fusion theory and Ensemble learning state that a combined indicator performs best when its basic components have little correlation between themselves, as indicators mutually overcome each others shortcomings. The scatter-plot and correlation matrix of the indicators is presented, respectively in Figures 11 and 12.

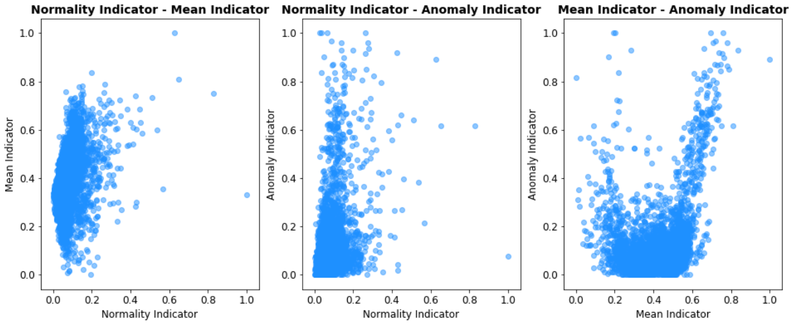


Figure 11. Scatter-plot of each pair combination of basic indicator.

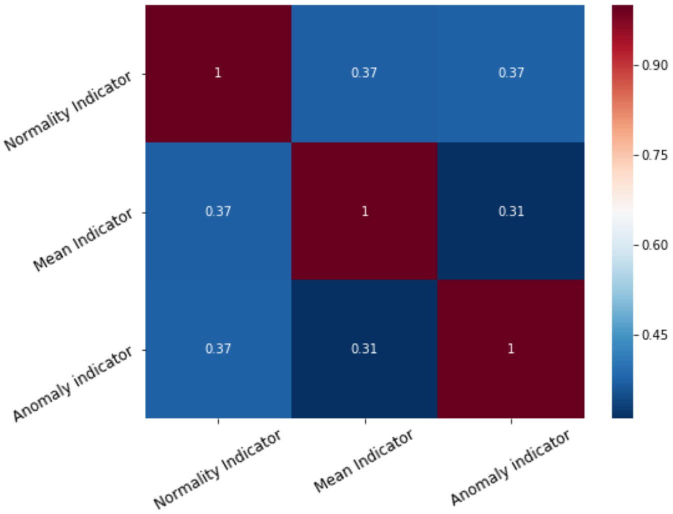


Figure 12. Correlation matrix of the base indicators.

The correlation coefficient of the indicators is never greater than 0.4. The amount of overlapped, redundant information is small, thus making their combination beneficial for overall predictive performances. Whenever additional indicators are added their correlation with the existing predictors should be checked. If two indicators are too similar, then only one should be used and the other may be discarded.

5.4. Failure Anticipation

Predictions, to be useful, must anticipate a failure by a sufficiently large amount of time, giving to the wind farm operator the possibility to organize the substitution of the broken component and adjust turbines production not to incur in fines due to missed production.

The verification of the anticipation margin is made observing a heatmap representation of the value of the combined indicator for the two analyzed wind farm. The combined indicator for wind farm 1 is shown in Figure 13. Two failures occurred and both of them are preceded by various weeks of high scores of the fault indicator value. A minimum of one month of anticipation of the main bearing failure is ensured.

Figure 14 presents results for wind farm 2. Both failures are correctly predicted with a safe margin of time allowing maintenance to be timely organized. Both heatmaps show turbines with high values of the combined indicator, without recorded maintenance interventions. This can be caused by concurring failures in other components or different operating conditions with respect to the rest of the wind farm. That being said, the ratio between false positives and true positives indicates that the proposed methodology offer a valid solution to automatize turbine reviews.

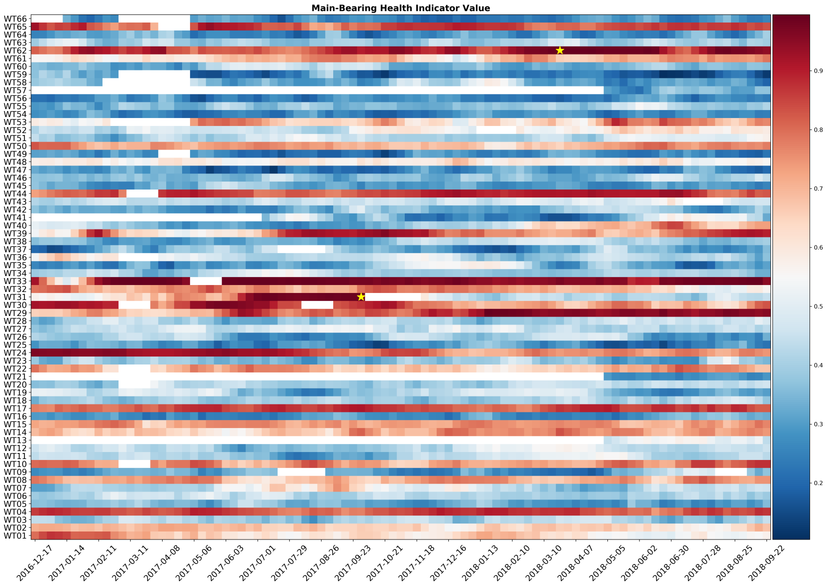


Figure 13. Heatmap of the combined main bearing health status indicator for wind farm 1. Failures are represented by a yellow star.

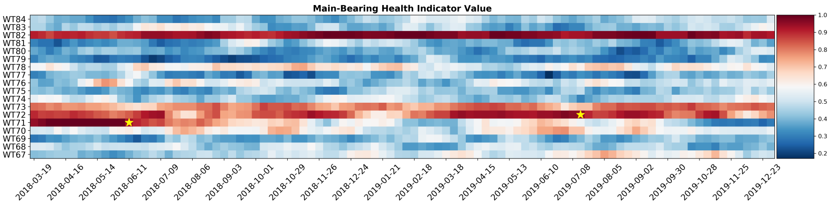


Figure 14. Heatmap of the combined main bearing health status indicator for wind farm 2. Failures are repersented by a yellow star.

6. Discussion

The proposed solution is characterized by an increased complexity of the decision process, when compared to Signal Trending or Normal Behavior Modeling techniques, as the information of multiple indicators is considered. Choosing a complete and significant set of indicators might be challenging, that being said, the presented results prove that it is a beneficial choice.

This strategy is highly modular, new indicators tailored to capture different behaviors of the data or utilizing other data streams can be easily incorporated into the decision process, once their complementarity to the already included indicators is verified. The use of multiple indicators based on detection of specific patterns in the data provides a more explainable interpretation of the behavior of the turbine with respect to complicated solutions processing data in a unique algorithm, often based on black-box structures.

The chosen indicators worked especially well for the detection of main bearing failure. As presented in Section 5.2, it is possible to set a high value of decision threshold without undermining failure detection. This means that wind farm operator do not require to check too many turbines to be sure to anticipate failures, a small number of revisions are necessary, following this strategy, and most of them will result in the discover of defects.

While more complex, the use of various indicators, proved especially beneficial in terms of elimination of FPs, as clearly shown in Section 5.3. Precision and F1 score greatly take advantage of the use of multiple indicators. In the first wind farm precision and F1 score almost doubled their values. The second wind farm benefits in a lower measure of the merging process, but significant improvements are observed.

Another remarkable characteristic of this approach is the ability to reliably anticipate failures, as debated in Section 5.4. It is critical to guarantee a margin of anticipation for main bearing failures, as the logistic is not trivial and a maintenance intervention cannot be arranged on a short-notice. As it is shown, the predictive methodology anticipated all four events by at least one month. Wind farm operators are then put in condition to adapt their production schedule and avoid losses due to unexpected and critical failures of main bearings.

Ultimately, the decision to avoid supervised learning solutions that require the time-consuming phase of data labeling helped to decrease greatly setup times of this architecture, repaying the additional time required to implement a set of multiple indicators and a merging strategy to aggregate their results.

7. Conclusions

This paper proposes a novel and innovative predictive maintenance solution based on Ensemble Learning using SCADA data, for wind turbine farms. The main characteristics of this solution can be summarized in three key-points:

- Unsupervised algorithms;
- Interpretable results;
- Combination of various indicators into a more reliable one via Ensemble Learning.

The time to pre-process and train algorithms is greatly reduced, as labeling of operating data into healthy and faulty conditions is not required. Incidentally, this techniques also has more flexible requirements, work orders are not necessary as they are used for evaluation purposes only. The presented algorithm only requires SCADA data to be put into production.

The indicators are designed on specific failure patterns, that are easy to interpret (drift in temperatures, changes in the relation of key variables...). The presented methodology has been rigorously tested on two year worthy of data from two onshore wind farms, for a total of 84 turbines.

Results proved that the combination of multiple indicators into a single predictor grants substantial improvements in performances, reaching an average accuracy of 95.1%, precision of 24.5% and F1 score of 38.5%. The sensitivity to key parameters as the threshold that discriminate faulty turbines from normal ones is studied, suggesting that high threshold values leads to good results, as the chosen indicators are able to isolate faulty from healthy turbines. The anticipation of failure, in all four events analyzed, is no less than one month giving wind farm operators time to organize logistics and minimize losses related to downtime.

Future researches may design additional indicators, as well as define tuning strategies of the decision threshold, incorporating maintenance costs and savings for early fault detection and optimize economic benefit of the predictive strategy. If vibration or acoustics data is available, new indicators could be designed and integrated to improve performances. It has to be noticed that we have been able to test this methodology on main bearing failures only, due to the limitations of the dataset at hand. Other turbine systems, such as gearbox and generator bearings or pitch actuators could have different failure signatures, thus other indicators might be needed and adjustments to the presented methodology required. The application of this strategy to monitoring of other components is a line of research that we warmly recommend to readers.

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

Funding: This research was funded by Centro para el Desarollo Tecnológico Industrial, grant number CDTI-IDI 20191294 and Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), grant number DOCTORADO AGAUR-2017-DI 004.

Acknowledgments: The authors want to acknowledge the support of the other colleagues at SMARTIVE for their support and technical advice in the design of this experiment and the analysis of the results.

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

Abbreviations

The following abbreviations are used in this manuscript:

AE	Autoencoder
csv	Comma Separated Value
EE	Elliptical Envelope
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
IF	Isolation Forest
LCOE	Levelized Cost Of Electricity
NBM	Normal Behavior Modeling
O&M	Operation and Maintenance
OCSVM	One Class Support Vector Machines
RMSE	Root Mean Squared Error
SCADA	Supervisory Control Furthermore, Data Acquistion
SOM	Self Organizing Map
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

References

1. Deployment, Investment, Technology, Grid Integration and Socio-Economic Aspects. 2019. Available online: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2019/Oct/IRENA_Future_of_wind_2019.pdf (accessed on 10 January 2021).
2. Wind Energy in Europe in 2019—Trends and Statistics. Available online: https://windeurope.org/data-and-analysis/product/wind-energy-in-europe-in-2019-trends-and-statistics/ (accessed on 13 January 2021).
3. Crabtree, C.J.; Zappalá, D.; Hogg, S.I. Wind energy: UK experiences and offshore operational challenges. *Proc. Inst. Mech. Eng. Part J. Power Energy* **2015**, *229*, 727–746. [CrossRef]
4. Hart, E.; Turnbull, A.; Feuchtwang, J.; McMillan, D.; Golysheva, E.; Elliott, R. Wind turbine main-bearing loading and wind field characteristics. *Wind Energy* **2019**, *22*, 1534–1547. [CrossRef]

5. Ahmed, M.A.; Kim, Y.C. Hierarchical communication network architectures for offshore wind power farms. *Energies* **2014**, *7*, 3420–3437. [CrossRef]
6. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring–a review. *IET Renew. Power Gener.* **2016**, *11*, 382–394. [CrossRef]
7. Maldonado-Correa, J.; Martín-Martínez, S.; Artigao, E.; Gómez-Lázaro, E. Using SCADA Data for Wind Turbine Condition Monitoring: A Systematic Literature Review. *Energies* **2020**, *13*, 3132. [CrossRef]
8. Hart, E.; Clarke, B.; Nicholas, G.; Kazemi Amiri, A.; Stirling, J.; Carroll, J.; Dwyer-Joyce, R.; McDonald, A.; Long, H. A review of wind turbine main bearings: Design, operation, modelling, damage mechanisms and fault detection. *Wind Energy Sci.* **2020**, *5*, 105–124. [CrossRef]
9. Guo, P.; Infield, D.; Yang, X. Wind turbine generator condition-monitoring using temperature trend analysis. *IEEE Trans. Sustain. Energy* **2011**, *3*, 124–133. [CrossRef]
10. Qiu, Y.; Feng, Y.; Sun, J.; Zhang, W.; Infield, D. Applying thermophysics for wind turbine drivetrain fault diagnosis using SCADA data. *IET Renew. Power Gener.* **2016**, *10*, 661–668. [CrossRef]
11. Tonks, O.; Wang, Q. The detection of wind turbine shaft misalignment using temperature monitoring. *CIRP J. Manuf. Sci. Technol.* **2017**, *17*, 71–79. [CrossRef]
12. Cambron, P.; Tahan, A.; Masson, C.; Pelletier, F. Bearing temperature monitoring of a Wind Turbine using physics-based model. *J. Qual. Maint. Eng.* **2017**. [CrossRef]
13. Sun, P.; Li, J.; Wang, C.; Lei, X. A generalized model for wind turbine anomaly identification based on SCADA data. *Appl. Energy* **2016**, *168*, 550–567. [CrossRef]
14. Artigao, E.; Koukoura, S.; Honrubia-Escribano, A.; Carroll, J.; McDonald, A.; Gómez-Lázaro, E. Current signature and vibration analyses to diagnose an in-service wind turbine drive train. *Energies* **2018**, *11*, 960. [CrossRef]
15. Siegel, D.; Zhao, W.; Lapira, E.; AbuAli, M.; Lee, J. A comparative study on vibration-based condition monitoring algorithms for wind turbine drive trains. *Wind Energy* **2014**, *17*, 695–714. [CrossRef]
16. Soua, S.; Van Lieshout, P.; Perera, A.; Gan, T.H.; Bridge, B. Determination of the combined vibrational and acoustic emission signature of a wind turbine gearbox and generator shaft in service as a pre-requisite for effective condition monitoring. *Renew. Energy* **2013**, *51*, 175–181. [CrossRef]
17. Ferrando Chacon, J.L.; Andicoberry, E.A.; Kappatos, V.; Papaelias, M.; Selcuk, C.; Gan, T.H. An experimental study on the applicability of acoustic emission for wind turbine gearbox health diagnosis. *J. Low Freq. Noise Vib. Act. Control.* **2016**, *35*, 64–76. [CrossRef]
18. Inturi, V.; Sabareesh, G.; Supradeepan, K.; Penumakala, P. Integrated condition monitoring scheme for bearing fault diagnosis of a wind turbine gearbox. *J. Vib. Control* **2019**, *25*, 1852–1865. [CrossRef]
19. Astolfi, D.; Castellani, F.; Terzi, L. Fault prevention and diagnosis through SCADA temperature data analysis of an onshore wind farm. *Diagnostyka* **2014**, *15*, 71–78.
20. Cambron, P.; Masson, C.; Tahan, A.; Pelletier, F. Control chart monitoring of wind turbine generators using the statistical inertia of a wind farm average. *Renew. Energy* **2018**, *116*, 88–98. [CrossRef]
21. Yang, W.; Court, R.; Jiang, J. Wind turbine condition monitoring by the approach of SCADA data analysis. *Renew. Energy* **2013**, *53*, 365–376. [CrossRef]
22. Feng, Y.; Qiu, Y.; Crabtree, C.J.; Long, H.; Tavner, P.J. Monitoring wind turbine gearboxes. *Wind Energy* **2013**, *16*, 728–740. [CrossRef]
23. Li, Y.; Wu, Z. A condition monitoring approach of multi-turbine based on VAR model at farm level. *Renew. Energy* **2020**, *166*, 66–80. [CrossRef]
24. Schlechtingen, M.; Santos, I.F. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mech. Syst. Signal Process.* **2011**, *25*, 1849–1875. [CrossRef]
25. Marti-Puig, P.; Blanco-M, A.; Serra-Serra, M.; Solé-Casals, J. Wind Turbine Prognosis Models Based on SCADA Data and Extreme Learning Machines. *Appl. Sci.* **2021**, *11*, 590. [CrossRef]
26. Zhang, Z.Y.; Wang, K.S. Wind turbine fault detection based on SCADA data analysis using ANN. *Adv. Manuf.* **2014**, *2*, 70–78. [CrossRef]
27. Bangalore, P.; Tjernberg, L.B. An artificial neural network approach for early fault detection of gearbox bearings. *IEEE Trans. Smart Grid* **2015**, *6*, 980–987. [CrossRef]
28. Zhao, H.; Liu, H.; Hu, W.; Yan, X. Anomaly detection and fault analysis of wind turbine components based on deep learning network. *Renew. Energy* **2018**, *127*, 825–834. [CrossRef]
29. Yang, W.; Liu, C.; Jiang, D. An unsupervised spatiotemporal graphical modeling approach for wind turbine condition monitoring. *Renew. Energy* **2018**, *127*, 230–241. [CrossRef]
30. Chen, P.; Li, Y.; Wang, K.; Zuo, M.J.; Heyns, P.S.; Baggeröhr, S. A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks. *Measurement* **2020**, *167*, 108234. [CrossRef]
31. Blanco-M, A.; Gibert, K.; Marti-Puig, P.; Cusidó, J.; Solé-Casals, J. Identifying health status of wind turbines by using self organizing maps and interpretation-oriented post-processing tools. *Energies* **2018**, *11*, 723. [CrossRef]

32. Du, M.; Ma, S.; He, Q. A SCADA data based anomaly detection method for wind turbines. In Proceedings of the 2016 China International Conference on Electricity Distribution (CICED), Xi'an, China, 10–13 August 2016; 2016; pp. 1–6.
33. McKinnon, C.; Carroll, J.; McDonald, A.; Koukoura, S.; Infield, D.; Soraghan, C. Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data. *Energies* **2020**, *13*, 5152. [CrossRef]
34. Purarjomandlangrudi, A.; Ghapanchi, A.H.; Esmalifalak, M. A data mining approach for fault diagnosis: An application of anomaly detection algorithm. *Measurement* **2014**, *55*, 343–352. [CrossRef]
35. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; 2008; pp. 413–422.
36. Koren, Y. The bellkor solution to the netflix grand prize. *Netflix Prize. Doc.* **2009**, *81*, 1–10.
37. Wu, Z.; Lin, W.; Ji, Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access* **2018**, *6*, 8394–8402. [CrossRef]
38. Wang, G.; Jia, R.; Liu, J.; Zhang, H. A hybrid wind power forecasting approach based on Bayesian model averaging and ensemble learning. *Renew. Energy* **2020**, *145*, 2426–2434. [CrossRef]
39. Lee, J.; Wang, W.; Harrou, F.; Sun, Y. Wind power prediction using ensemble learning-based models. *IEEE Access* **2020**, *8*, 61517–61527. [CrossRef]
40. Liu, Y.; Cheng, H.; Kong, X.; Wang, Q.; Cui, H. Intelligent wind turbine blade icing detection using supervisory control and data acquisition data and ensemble deep learning. *Energy Sci. Eng.* **2019**, *7*, 2633–2645. [CrossRef]
41. Turnbull, A.; Carroll, J.; McDonald, A. Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure. *Wind Energy* **2020**. [CrossRef]
42. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, 1189–1232. [CrossRef]
43. Marti-Puig, P.; Blanco-M, A.; Cárdenas, J.J.; Cusidó, J.; Solé-Casals, J. Effects of the pre-processing algorithms in fault diagnosis of wind turbines. *Environ. Model. Softw.* **2018**, *110*, 119–128. [CrossRef]
44. Lebranchu, A.; Charbonnier, S.; Bérenguer, C.; Prévost, F. A combined mono-and multi-turbine approach for fault indicator synthesis and wind turbine monitoring using SCADA data. *ISA Trans.* **2019**, *87*, 272–281. [CrossRef]
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Gradient Tree Boosting. Available online: https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting (accessed on 5 January 2021).
47. Sklearn.Ensemble.IsolationForest. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html (accessed on 18 January 2021).

4.3 Improved Ensemble Learning for Wind Turbine Main Bearing

The second research paper is also based on an ensemble learning architecture but it uses more complex and less interpretable algorithms. An anomaly detection algorithm —isolation forest— is combined to a Normal Behavior Model implemented through a neural network. Compared to the previous contribution, interpretability is reduced favoring predicting performances instead.

The combination of these two indicators led to very positive results, specifically the number of false alarms was greatly reduced. As discussed in the paper, these results have been obtained due to the limited correlation of the indicators.

Contributions

The main novelties introduced by the research are:

- **Implement** an unsupervised approach, no labeled data is needed to train the algorithms.
- Ensemble learning is used to **combine** two different predicting algorithms capturing distinct patterns in the data.
- **Different** training and prediction **window sizes** can be set.
- Fault predictions anticipated the replacement of the component by several months, allowing maintainers to **optimize logistics and minimize costs**.
- Results are obtained using **a large set of real data**.

Additionally, it is shown how the ensemble framework is a very modular solution that can be easily extended to accommodate new algorithms. Moreover, the balance between interpretability and reliability of predictions can be controlled through the choice of the base learners.

applied sciences

Article

Improved Ensemble Learning for Wind Turbine Main Bearing Fault Diagnosis

Mattia Beretta [1,2], Yolanda Vidal [3,4], Jose Sepulveda [2], Olga Porro [2] and Jordi Cusidó [2,5,*]

[1] Unitat Transversal de Gestió de l'Àmbit de Camins (UTGAC), Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain; mattia.beretta@upc.edu

[2] SMARTIVE S.L., 08204 Sabadell, Spain; jose.sepulveda@smartive.eu (J.S.); olga.porro@smartive.eu (O.P.)

[3] Control, Modeling, Identification and Applications (CoDAlab), Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Campus Diagonal-Besós (CDB), Universitat Politècnica de Catalunya (UPC), Eduard Maristany, 16, 08019 Barcelona, Spain; yolanda.vidal@upc.edu

[4] Institute of Mathematics (IMTech), Universitat Politècnica de Catalunya (UPC), Pau Gargallo 14, 08028 Barcelona, Spain

[5] Enginyeria de Projectes i de la Construcció EPC, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain

* Correspondence: jordi.cusido@smartive.eu; Tel.: +34-620-602-495

Abstract: The goal of this paper is to develop, implement, and validate a methodology for wind turbines' main bearing fault prediction based on an ensemble of an artificial neural network (normality model designed at turbine level) and an isolation forest (anomaly detection model designed at wind park level) algorithms trained only on SCADA data. The normal behavior and the anomalous samples of the wind turbines are identified and several interpretable indicators are proposed based on the predictions of these algorithms, to provide the wind park operators with understandable information with enough time to plan operations ahead and avoid unexpected costs. The stated methodology is validated in a real underproduction wind park composed by 18 wind turbines.

Keywords: fault prognosis; WT; main bearing; normality model; real SCADA data

check for updates

Citation: Beretta, M.; Vidal, Y.; Sepulveda, J.; Porro, O.; Cusidó, J. Improved Ensemble Learning for Wind Turbine Main Bearing Fault Diagnosis. *Appl. Sci.* **2021**, *11*, 7523. https://doi.org/10.3390/app11167523

Academic Editor: Mohsen Soltani

Received: 6 July 2021
Accepted: 12 August 2021
Published: 17 August 2021

1. Introduction

Global demand for energy has achieved an unprecedented level with the growing world population and the rising industrialization in developing countries. However, globally, the largest amount of energy is obtained from fossil fuels, which is closely related to the rising levels of greenhouse gases emissions. Definitely, energy transition to renewable sources is the crux of the matter to fight climate change. As stated in [1] only 10% of the world's primary energy supply is already from renewable energies, but they are steadily growing. Among renewable energy sources, wind power has been the fastest growing in the recent decades, and in 2019 it was the leading source of new capacity in Europe, the US, and Canada, and the second largest in China [2].

It is noteworthy that wind energy levelized cost of energy (LCOE) fell 39% between 2010 and 2019 [3]. However, from this LCOE, the operation and maintenance (O&M) accounts for 28.5% in land-based wind projects, and up to 34% in offshore wind projects [4]. Energy production losses due to downtime (caused by O&M of the assets), together with the costs associated to the replacement of components can scale up to millions of Euros per year in any industrial size wind park. Thus, it is of paramount importance that the wind industry moves from corrective (repairing components after they break down) and preventive maintenance (scheduled at regular intervals) to predictive maintenance (scheduled as needed based on the asset condition). Predictive, or condition-based maintenance (CBM), provides operators with an advanced warning before the actual fault occurs, allowing them to plan ahead and schedule repairs to coincide with weather or production windows to reduce costs and turbine downtime. CBM is an extensive area of research in a wide variety of applications, such as smart manufacturing [5]. However, its application to complex systems, such as wind turbines (WTs) that work under different and varying operating and environmental conditions remains an open challenge. Furthermore, the latest CBM developments tend to use expensive specifically tailored sensors for condition monitoring (CM), which is not economically viable for turbines already under operation and even less in case they are close to reach the end of their lifespan.

The life expectancy of wind parks from the late 1990s and early 2000s surge of wind power is about to be completed. A decision on lifetime extension is complicated, but it is clear that end-of-life solutions will develop a significant market over the next five years [6]. In particular, CBM data-driven methodologies based on already available supervisory control and data acquisition (SCADA) data are a promising cost-effective solution. These SCADA data are highly variable because of the changing operational conditions, they have a low sampling time (10 min average value), are not standardized, and have not been initially designed for the particular purpose of CM but for control purposes. Therefore, it is a hard challenge to contribute CBM strategies based solely on these data [7]. However, this is an active research area, as shown by recent publications. On the one hand, much of the references found in the literature work with simulated SCADA (e.g., [8]) that can be obtained using open source simulation software and, rarely, real data as these are proprietary data from the wind park operators and are not easily available. On the other hand, when dealing with real data, normally, only one or two WTs are tested. For example, [9] where support vector machines are used to predict WT faults, and [10], where different machine learning classifiers are compared to predict generator faults. These references use real WT SCADA data from one and two WTs, respectively, and are based on supervised approaches that require historical fault data to be constructed. This is an important drawback, as obtaining labeled datasets from operational data is typically hard, it is exposed to errors, and leads to a highly unbalanced dataset. Additionally, the methodology can not be applied straightforward to wind parks where the fault of interest did not occur in the past. Thus, despite the promising performance of supervised methods, unsupervised approaches are preferred for SCADA predictive maintenance [11]. The recent works related to unsupervised SCADA based predictive maintenance can be mainly grouped in one of the following three categories: signal trending, normality models, and anomaly detection models. In the following paragraphs, a brief review of each one of these categories is given.

The approaches in the signal trending category study changes and trends in the SCADA time series. Some relevant works include [12], where monitoring of WT generators using the statistical inertia of a wind park average is proposed; and [13], where the difference between the SCADA data of each turbine with the median of the rest of turbines in the wind park is used to establish a condition vector to later apply vector autoregression Hotelling and vector autoregression multivariate exponentially weighted moving average to locate the faulty turbine. The main advantages of signal trending methods are its simplicity and interpretability. However, most of these methods can not capture complex relations among different variables, thus machine and deep learning models are needed to move beyond signal trending solutions.

Regarding data-based normality models, also called normal behavior models, their objective is to learn the relation between a set of input variables and a target variable under normal operation. Then, the difference between predictions and real measurements (for the target variable) is used to detect abnormal behavior. This normality models are well established in unsupervised SCADA based predictive maintenance, but the general trend is to use the power curve (relation between the wind intensity and the extracted power) as the target variable. It is noteworthy the work of [14] where main working parameters (e.g., the rotor speed, and the blade pitch) are used as input variables and the power is employed as the target to construct the normality model. The main advantage of these methods is their capability to learn complex relations between different sensors. On the other hand, its major drawback is a lack of interpretability.

The main bearing is the object of this analysis due to its central role in the drive train assemble. It connects the rotor to the gearbox, and it has to withstand the large torque generated by the rotor, making it prone to problems. Failure rates of 30% are reported for single main bearing and 15% for double main bearing calculated over a 20 year lifetime [15]. Within the typical wear mechanisms affecting the main bearing, micro-pitting, spalling, and smearing can be listed [16]. Main bearing failures can be anticipated through vibration analysis, but also using predictive models based on SCADA data and analysis of temperature signals, as reported in [17].

In this work, a normality model at WT level is selected, based in [18], where the target variable is selected to be the closest sensor to the component under study. In particular, as the main bearing is the component to be monitored, the main shaft temperature is used as target variable. In this work, the component under study is the main bearing because, as stated by the European Academy of Wind Energy (EAWE) [16], the wind industry has identified main bearing failures as a critical issue in terms of increasing WT reliability and availability, as they lead to major repairs with high replacement costs and long downtime periods.

The methods in the category of anomaly detection models, also called outlier detection, seek to identify rare samples which raise suspicion by differing significantly from the majority of the data. A significant work is [19] where a comparison of three anomaly detection techniques (one-class support vector machine, isolation forest, and elliptical envelope) for WT CBM using SCADA data is realized. Isolation forest has some advantages over other approaches as it does not require a normal operation dataset, and it requires less training data than other deep learning strategies. In this work, an anomaly detection model at wind park level is proposed based on the isolation forest methodology.

Considering all the aforementioned references, in this work, a CBM strategy based on SCADA data is stated with a five-fold contribution: (i) It is an unsupervised approach, thus there is no need that the specific studied fault happened in the past to train the proposed models; (ii) It is an ensemble that combines the benefits of a WT normal behavior model with a wind park anomaly detection model; (iii) It combines different training and prediction window sizes; (iv) Fault prediction is accomplished months in advance prior to the fault, giving enough time to operators to plan ahead and schedule repairs; and (v) The validity and performance of the proposed methodology is demonstrated (tested) on a dataset covering two years and a half of operation from a real underproduction wind park composed by 18 WTs.

The rest of the article is organized as follows. Section 2 presents the available SCADA data and work order logs used in this work. Next, Section 3 states the proposed ensemble methodology for WT main bearing fault diagnosis, including a comprehensive description of the single models and their indicators. Section 4 presents the results and their discussion to interpret and describe the significance of the ensemble method in comparison to the single models by themselves. Finally, in Section 5, conclusions are drawn, and future work is proposed.

2. Data

Two information sources are used in this research: (i) SCADA operating data and (ii) work order logs. The first one is used for modeling the behavior of the turbine, whereas the second one is used uniquely to validate the results of the algorithms. Both SCADA and work order logs are typically available to wind park owners, without the need of installing new sensors nor change operating routines. Albeit, the format and content of the work order logs, being less structured and not standardized, may vary significantly from one wind park maintainer to another.

2.1. SCADA Operating Data

SCADA is made off a vast net of sensors, monitoring the state of the main components of a turbine, as well as the environmental conditions. Operating data are recorded at high frequency and successively down-sampled to reduce network usage and storage costs.

The resolution of SCADA data is usually of 10 min. To reduce information loss, due to down-sampling, not only the mean value across the aggregation period is stored, but also the standard deviation, minimum, and maximum values. An example of a typical SCADA dataset is provided in Table 1.

Table 1. Sample of SCADA data for a limited selection of turbines and signals. Each signal is stored completed with its minimum (*min*), maximum (*max*), mean (*mean*), and standard deviation (*std*) values. The frequency of the data is 10 min.

		Wind Speed [m/s]				**Main Bearing Temperature [°C]**			
		min	***max***	***std***	***mean***	***min***	***max***	***std***	***mean***
WT	**Timestamp**								
WT01	2018-01-01 00:00:00	3.8	11.8	1.369	8.346	33.0	33.0	0.000	33.000
WT02	2018-01-01 00:00:00	3.8	11.5	1.234	8.065	32.0	33.0	0.221	32.051
WT03	2018-01-01 00:00:00	5.6	11.5	0.976	8.283	29.0	30.0	0.500	29.505
WT01	2018-01-01 00:10:00	3.7	11.1	1.323	7.794	33.0	33.0	0.000	33.000
WT02	2018-01-01 00:10:00	3.5	10.3	1.152	7.178	32.0	33.0	0.386	32.182
WT03	2018-01-01 00:10:00	4.6	10.9	0.984	7.858	29.0	30.0	0.499	29.532

Operating data from a European onshore wind park, situated in Poland, is used in this research. A total of 18 turbines with nominal power of 2.3 MW is available. Approximately three and a half years of data are analyzed, starting from the beginning of 2017 to mid-2020. Data from 2017 are used to train the normality model, and the remaining two and a half years are used to test predictions. Having more than one year of data allows to control seasonal variations in environment temperature and wind-speed.

2.2. Work Orders

Wind park owners commonly keep track of the ordinary and extraordinary maintenance interventions required by the turbines. This information is typically stored in text files, where a description of the intervention is reported with a timestamp of the date in which it occurred. Sometimes detailed information, such as the material required for the intervention and other details, is provided. In this research, work order logs are used to assess the validity of the main-bearing status predictions, and it is not fed as input to the algorithms. An example of the work order logs is presented in Table 2.

Table 2. Sample of work order data.

WT	**Timestamp**	**Component**	**Comments**
WT11	2017-04-28 11:08:00	Gearbox Bearing	Gearbox Bearing repair
WT06	2018-06-11 08:40:00	Main Bearing	Replacing Main Bearing and Main Shaft
WT07	2019-07-17 07:40:00	Main Bearing	Inspection required, condition-based.
WT03	2020-02-25 09:10:00	Main Bearing	Main Bearing inspection, due to rate of worn

3. Methodology

3.1. Normal Behavior Model

In this subsection, the selected normal behavior model based on an artificial neural network (ANN) is comprehensively described. Note that this model is designed at a turbine level, thus each WT in the park will have its associated normality model trained with only its own historical SCADA data. First, the data preprocess is detailed to deal with out-of-range values, missing data, and sensors with different magnitudes. Second, a one-year time window is selected to define the training dataset including all operating conditions of the WT. This will ensure that the detected anomalies are not just a change in seasonality and will allow the methodology to be used across all regions of operation of the WT. Third, the ANN set-up is detailed, including a brief explanation of the optimization and regularization methods, as well as the selection of the ANN structure. Finally, a specific purposely build indicator for the normal behavior model is stated.

3.1.1. Data Preprocess

First, variable selection to determine a set of variables that will provide the best fit for the model so that accurate predictions can be made is needed. In this work, the temperatures of the components located close to the main bearing are selected as variables of the normality model together with the ambient temperature, as it affects the temperatures of all subsystems. Additionally, the generated power and rotor speed provide information about the region of operation of the WT and, thus, they are also used as variables. In summary, the selected variables are shown in Table 3 where also the specific ranges of realistic values for each sensor are listed.

Table 3. Selected SCADA variables to develop the normal behavior model, its description, range of possible values, and units. All of them are related to the mean value over a 10-min period.

Variable	Description	Range	Units
Power	Generated real power	[0, 2000]	kW
AmbientTemp	Ambient temperature	[−5, 40]	°C
BearingCSTemp	Bearing coupling side temperature	[0, 120]	°C
BearingNCSTemp	Bearing non-coupling side temperature	[0, 120]	°C
LowSpeedShaftTemp	Low-speed shaft temperature	[0, 120]	°C
GeneratorTemp	Generator temperature	[0, 175]	°C
GearboxTemp	Gearbox temperature	[0, 120]	°C
RotorSpeed	Rotor speed	[0, 50]	rpm

Second, for each variable out-of-range values are deleted as they are associated with sensor measurement errors. Furthermore, data imputation of missing values (and deleted values of the prior step) is carried out through piecewise cubic Hermite interpolating polynomials. As it has a local smoothing property, this strategy produces more stable estimates compared to other standard approaches used for data imputation [20]. Notice that the nearest value before or after the missing values is used at the beginning and end of the dataset, respectively.

Finally, as data from various variables have varying magnitudes, the max–min normalization is used to scale the dataset.

3.1.2. Train and Test Sets

The aim of the normality model is that it is capable to cope with the various operational and environmental conditions that the WT will face, see [18]. Thus, the train and test datasets include data from all working conditions: different wind speed regions and their associated regions of operation of the WT (start up, maximize power, and limit wind power to avoid exceeding the safe electrical and mechanical loads), different year seasons, curtailment, etc. Therefore, it is noteworthy that in this work there is no filtering of the data based on specific regions of operation or seasonality. The available SCADA data, for the normality model, are divided as follows: data from 2017 are used to train the model (thus, the training dataset contains a whole year of data), and the remaining two and a half years are used to test predictions. Note that there is no validation set because Bayesian regularization is used to train the ANN.

Finally, recall that a customized normality model will be built for each wind turbine in the park. This could raise some concerns related to its computational cost. However, note that after the model is trained on one year data, this model is used for two-year predictions ahead, as will be shown in the results section. Thus, the computational cost is low in comparison to the use of a training rolling window of observations preceding the target forecast that must be retrained at each window shift of the data.

3.1.3. ANN Set Up

The ANN model structure is proposed in this section and is based on the eight selected variables, shown in Table 3. The output of the ANN is the temperature of the low-speed shaft (variable of interest) at time t, and the inputs are the following ones:

y_1: generated real power at time $t-1$,
y_2: generated real power at time t,
y_3: ambient temperature at time $t-1$,
y_4: ambient temperature at time t,
y_5: bearing coupling side temperature at time $t-1$,
y_6: bearing coupling side temperature at time t,
y_7: bearing non-coupling side temperature at time $t-1$,
y_8: bearing non-coupling side temperature at time t,
y_9: generator temperature at time $t-1$,
y_{10}: generator temperature at time t,
y_{11}: gearbox temperature at time $t-1$,
y_{12}: gearbox temperature at time t,
y_{13}: rotor speed at time $t-1$,
y_{14}: rotor speed at time t.

Thus, referring to the structure of the ANN, there are 14 inputs noted as y_i, $i = 1, \cdots, 14$, there is 1 output noted as $\widehat{y}$, and a hidden layer comprising 72 neurons. Figure 1 shows the ANN architecture.

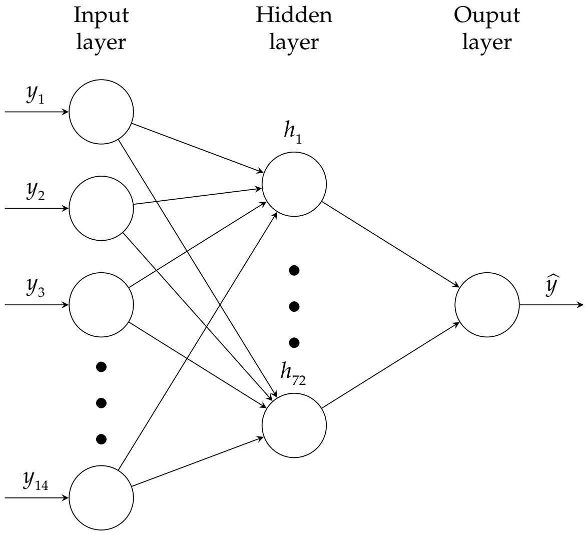


Figure 1. ANN proposed architecture. There are 14 inputs. The hidden layer is set to 72 nodes. The output layer is the estimated temperature of the low-speed shaft at time t.

The ANN is trained with the Levenberg–Marquardt optimization method combined with Bayesian regularization (to enhance the generalization capability of the model) as they are able to obtain lower mean squared errors than any other algorithms for functional approximation problems, see [21,22]. Recall that the main purpose of the WT normality model is to approximate precisely a function, namely the temperature of the low-speed shaft of that specific WT, under normal (healthy) condition. The Bayesian selection of the regularization parameters provides an optimal regularized solution, as well as insight into the effective number of parameters actually used by the ANN which is extremely useful to design the size of the network. In this work a value of 1058 effective number of parameters is obtained from a total of 1153 parameters in the proposed network (number of weights and biases), thus the complexity of the ANN is appropriate for the used training dataset. A comprehensive description of the Levenberg–Marquardt with Bayesian regularization method can be found, for example, in [18,21,23].

Finally, note that the network uses rectified linear unit (ReLU) activation functions, and weights and biases initialization is performed using the Xavier initializer [24].

3.1.4. Normal Behavior Model Indicator

A fault indicator is needed that activates an alarm when samples not following the normality model are detected. Initially, such an indicator could be based on establishing a threshold in the residual between the estimated value (given by the ANN) and the real SCADA data. However, this will result in an unacceptably high number of false alarms because of solitary samples trespassing the threshold, rendering the method worthless. Therefore, it is critical to establish an indicator that takes into account the persistence of consecutive samples above a certain threshold.

First, a detection threshold is prescribed as $\mathsf{threshold} = \mu + 6\sigma$, based on the mean, μ, and standard deviation, σ, of the residuals over the training data.

Second, a weekly indicator is implemented as follows:

$$\mathsf{indicator}_{\mathsf{normality}} = \min\left(1, \frac{n_{over}}{504}\right). \tag{1}$$

where n_{over} denotes the number of samples that had a residual value greater than the threshold that week. Note that this indicator has a range between 0 and 1. When all samples are below the threshold, the indicator value is 0. When half or more of the samples in a week are above the threshold, the indicator value is 1.

3.2. Isolation Forest

A second method is used to analyze the status of the main bearing, namely an anomaly detection algorithm: isolation forest. The objective is to complement the prediction of the neural network normality model, using an algorithm that is able to analyze the wind park as a whole. First, data pre-processing is discussed, highlighting the differences to the processing required by the neural network. Second, the testing scheme for the isolation forest is presented; the main difference with the previous algorithm is the lack of a training period. Third, a brief explanation of isolation forest is provided together with a discussion on the selection of the main parameters for the algorithm. Finally, the post-processing of the anomaly detection results is detailed, showing how to construct an indicator that captures differences within turbines.

3.2.1. Data Preprocess

In analogy to the neural-network pre-processing, a reduced selection of variables is used. Main bearing temperature is chosen, as monitoring the status of this component is the objective of the research. Additionally, ambient temperature and rotating speed of the main shaft are selected as they determine information on the context and operating status of the main bearing. A summary table of the inputs is proposed in Table 4, completed with the ranges used to filter outliers and sensor measurement errors. No data imputation is performed for missing values, instead data are down-sampled from ten minutes to hourly resolution, reducing the amount of empty timestamps and cutting computation time. Being the main bearing a massive component the reduction in data frequency is mitigated by the large thermal inertia of the bearing, thus limiting information loss.

Table 4. Selected SCADA variables used to develop the anomaly detection model, its description, range of possible values, and units. All of them are related to the mean value over a 10-min period.

Variable	Description	Range	Units
AmbientTemp	Ambient temperature	[−5, 40]	°C
MainBearingTemp	Main bearing temperature	[0, 120]	°C
RotorSpeed	Rotor speed	[0, 50]	rpm

3.2.2. Testing Scheme

The way the isolation forest algorithm is used in this research, requires no selection of a set of normal conditions to adjust the parameters of the model. Instead, data are analyzed through a rolling window of 1 month width and instances lying at the border of the data distribution are marked as anomalies. The rolling window is shifted each time by one week, effectively sweeping the entire dataset. The choice of the window length is the fruit of a compromise between computation time and the ability to capture a complete set of operating conditions of the turbines.

3.2.3. Isolation Forest Setup

Isolation forest has been introduced by Liu et al. [25] as an efficient anomaly detection algorithm based on widely known decision trees. Unlike most model-based algorithms, no normal condition training set is required. Points lying at the edges of the data distribution are to be intended as anomalies. Having this definition of anomalous point, isolation forest uses binary search trees to recursively split data. Various fully developed trees, i.e., having single instance terminal nodes, are trained. Isolation of anomalies is promoted by using random partitions when separating data, as indicated in [25]. The anomaly score used to determine the likelihood of a point to be anomalous is based on the average path length required by the trained trees to isolate the cited point from the rest of the data. Equation (2) is the canonical definition for an isolation forest anomaly score.

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{2}$$

where $E(h(x))$ is the average of the path length $h(x)$ required to isolate the given point. The denominator $c(n)$ is the average path length of unsuccessful search in binary search tree. Given a distribution d, three cases are typical:

- s →1 $\implies$ x is anomalous
- s < 0.5 $\implies$ x is normal
- s ≈ 0.5 ∀ x in d $\implies$ no anomalies in d

In this research, the implementation of Isolation Forest from Sklearn, for Python programming language is utilized [26]. The following parameters are set to run the algorithm:

- number of estimators = 250
- contamination = 0.1
- max samples = 0.3

The number of estimators determines the quantity of decision trees of which the isolation forest is composed. More trees lead to more reliable estimations of the path length, at the cost of higher computational time. The contamination coefficient represents the expected amount of anomalies in the data. The points in the analyzed distribution are sorted by their anomaly score and the most anomalous ones, according to the contamination value, are selected. Finally, the maximum number of samples to use to train each estimator, i.e., each decision tree, is defined by the homonymous parameter.

3.2.4. Anomaly Detection Indicator

An anomaly indicator is defined to translate the output of the isolation forest into concrete information on the WT status. The discovered anomalies are assigned to the corresponding turbines, to define which turbines have a larger percentage of anomalous points with respect to the rest of the wind park. The indicator is calculated as the ratio between the anomalous points ($n_{anomalies}$) and the total amount of records (n_{total}) of a turbine for a given period of time, see Equation (3). The idea is that if no clear differences between turbines' behaviors are present, the anomaly indicator will be similar across the wind farm. On the other hand, if a turbine is subjected to different operating conditions, its percentage of anomalies over the total number of records will be higher if compared with other turbines.

$$\mathsf{indicator}_{\mathsf{anomaly}} = \frac{n_{anomalies}}{n_{total}}. \tag{3}$$

3.3. Ensemble

Ensemble strategies have been widely used in the literature both in power output prediction, as well as fault diagnosis [27–29].

The output of the previous algorithms is merged into a composed indicator, with the objective to produce better predictions by overcoming the limitations of the two methods. To build a valuable ensemble, it is important to choose base indicators that are complementary and not redundant, thus a key step is to verify that the correlation of the indicators is limited. The correlation of the indicators is addressed in Section 4.3. The normality indicator provides an individual analysis of the turbines, by comparing a historical model with current conditions, whereas the anomaly indicator is based on a comparison within the different turbines of the wind farm.

Ensemble Indicator

The ensemble strategy that is used in this research does not require training of an additional meta-learner, instead the normality indicator (1) and anomaly indicator (3) are combined through a rolling windowed sum. The pseudocode to compute the ensemble indicator is given in Algorithm 1. In particular, the first step is to rank the turbines, assigning the corresponding percentile in which they fall when compared to the indicator distribution of the whole wind farm. Possible ties between turbines are managed using a ranking scheme that assigns to the tying turbines the lowest rank of the group. Being the single model indicators calculated on a weekly basis, each turbine will be assigned a percentile rank per each week. The next step consists in the application of a rolling window sum. More specifically, a window of $p = 4$ weeks size is chosen. Applying a rolling window sum allows to highlight long-lasting changes, limiting the influence of isolated peaks. Finally, a decision threshold is applied to decide whether an alarm must be or not triggered.

Algorithm 1: Computation of the ensemble indicator

```
Result: $i_e$: indicator$_{\text{ensemble}}$.
Given $i_a$, $i_n$: indicator$_{\text{anomaly}}$, indicator$_{\text{normality}}$ respectively;
Set DT: decision threshold (e.g. 0.85);
Set $p$: rolling window size (e.g. 4);
STEP 1: Rank turbines w.r.t. wind farm;
for $w$ in [$w$, $w$+1, ..., $w$+$p$-1] do
    for $wt$ in turbines do
        $x_{[anomaly]wt,w}$ = percentile of $i_a$ for turbine $wt$ w.r.t. distribution of the entire windfarm during week $w$;
        $x_{[normality]wt,w}$ = percentile of $i_n$ for turbine $wt$ w.r.t. distribution of the entire windfarm during week $w$;
    end
end
STEP 2: Calculate the ensemble ranking during the period of observation for each turbine;
for $wt$ in turbines do
    $x_{[ensemble],wt} = \sum_{w=1}^{w+p-1} x_{[anomaly],wt,w} + x_{[normality],wt,w}$;
end
$x_{[max-ensemble]} = 2p$;
STEP 3: Calculate the ensemble indicator of a turbine;
$i_{e,wt} = \frac{x_{[ensemble]wt}}{x_{[max-ensemble]}}$;
STEP 4: Apply decision threshold;
if $i_{e,wt} > DT$ then
    Turbine requires maintenance;
else
    No maintenance required;
end
```

4. Results

4.1. Test Data

The proposed methodology is validated with real SCADA data from an underproduction wind park composed by 18 wind turbines. Recall that approximately three and a half years of data are analyzed, starting from the beginning of 2017 to mid-2020. Data from 2017 are used to train the normality model, and the remaining two and a half years are used to test predictions using the work order logs. Table 5 shows the main bearing failures during the test period based on the work order logs. Three WTs in the park suffer from the main bearing failure during the test period. Note that the work order log comments are different for each case, ranging from lubrication to component replacement.

Table 5. Main bearing failures reported in the work order logs.

WT	Timestamp	Component	Comments
WT03	2020 February 25 9:10:00	Main bearing	Main Bearing inspection, due to rate of worn
WT05	2018 June 11 8:40:00	Main bearing	Replacing Main Bearing and Main Shaft
WT06	2019 July 17 7:40:00	Main bearing	Inspection required, condition-based

4.2. Performance Metrics

To evaluate the performance of the proposed methodology, the results will be analyzed via a confusion matrix based on the criteria that is described next. When the alarm is triggered within a six-month window before the failure actually occurs, the result is reported as a true positive (TP). The method in this case provides sufficient advance notice of the failure, so that the wind park operators can organize the maintenance operation while minimizing costs. When the alarm is not triggered and the WT does not have any work order, the result is documented as a true negative (TN). In this case, the method is correctly establishing that the WT is under normal operation. When the alarm is not triggered, or it is done out of the six-month before failure window, and the WT suffers a main bearing failure, the result is reported as a false negative (FN). Finally, when the alarm is triggered but the WT does not have any work order, this result is informed as a false positive (FP).

Finally, the following performance metrics are used to analyze the results: accuracy, precision, recall, specificity, and F1 score. Their definitions are briefly recalled hereby,

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}},$$

$$\text{F1} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FN+FP})}.$$

4.3. Ensemble vs. Single Models

The different models that compound the ensemble should provide complementary information, see [17]. Thus, it is important to analyze the correlation between the indicators provided by each of the individual models: the anomaly detection model vs. the normality model. Figure 2 displays the isolation forest indicator vs. the neural network indicator for each WT in the park separately. The Pearson correlation coefficient and p-value for testing non-correlation are reported as r and p, respectively. It is shown that, in general, there is no correlation between the individual indicators (low values of the Pearson correlation coefficient are obtained), thus they are complementary in terms of their contained information. Note the particular case of WT05 where the Pearson correlation is about 0.78, which indicates that there is a moderate positive relationship between the variables, as the p-value in this case is almost zero (which indicates that the correlation coefficient is significant). The explanation of this special behavior is that WT05 suffers the only main bearing fault in the park that is correctly predicted by both indicators, thus in this case the indicators are correlated.

Figures 3–5 show the normal behavior model indicator, the anomaly detection indicator, and the ensemble indicator, respectively. In these figures, the blue stars indicate the occurrence of a main bearing fault (as recorded in the work order logs). Note that the normal behavior indicator shows high values in WT01 and WT18, which are healthy over the whole test set, but the anomaly detection indicator shows low values for these WTs. On the contrary, the anomaly indicator shows high values for WT16, which is healthy, but the normality indicator shows low values for this WT. The ensemble indicator adequately combines the information from the single indicators showing low indicator values for WT01, WT16, and WT18 as desired.

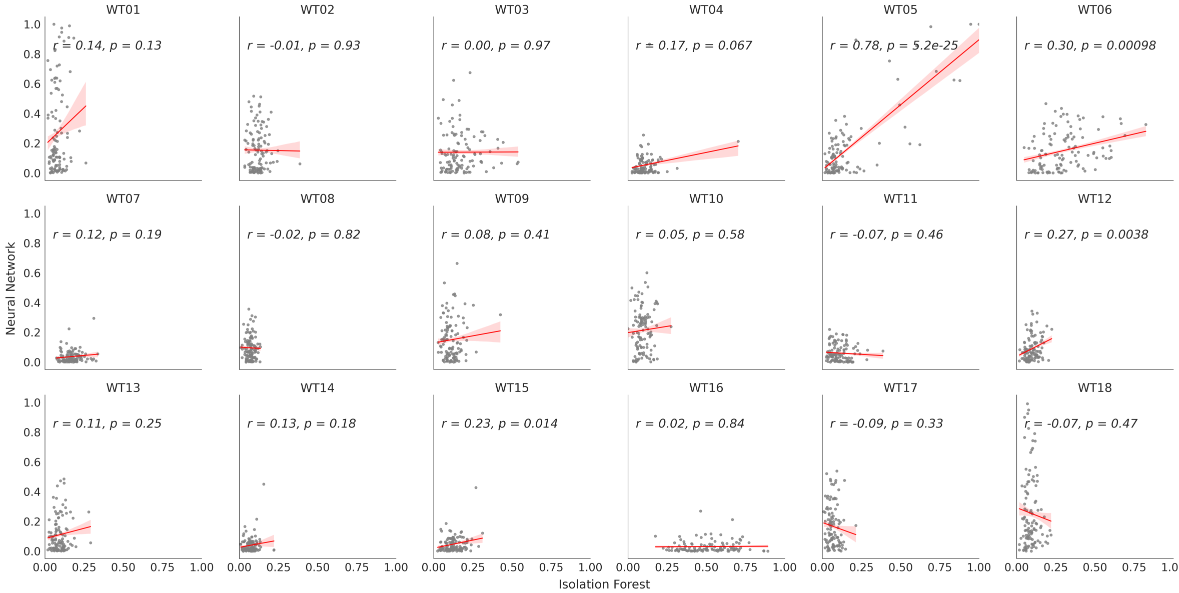


Figure 2. Scatter plot showing the isolation forest indicator vs. the neural network indicator independently for each WT. The linear regression model fit, that minimizes the squared error of the data, is plot in red. The shaded area represents the 68% confidence interval of the model. Pearson correlation coefficient and *p*-value for testing non-correlation are reported as *r* and *p*, respectively.

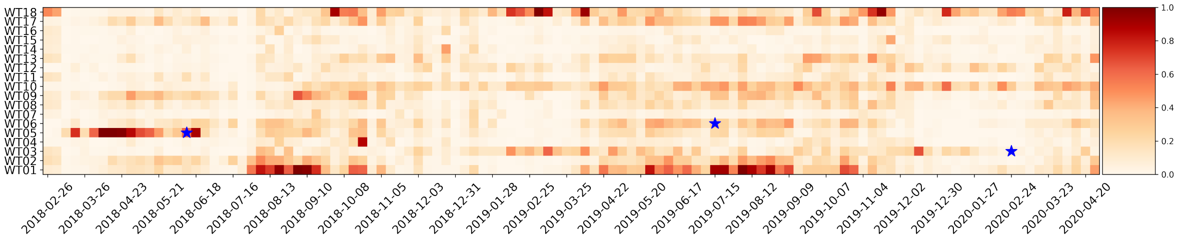


Figure 3. Main bearing health according to normal behavior model. Faults are marked with a blue star.

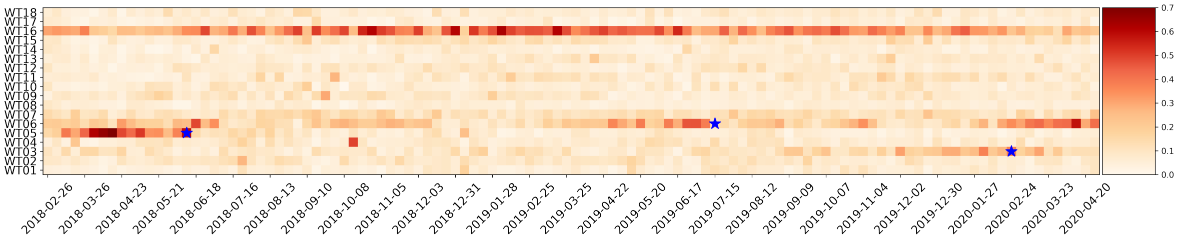


Figure 4. Main bearing health according to anomaly detection. Faults are marked with a blue star.

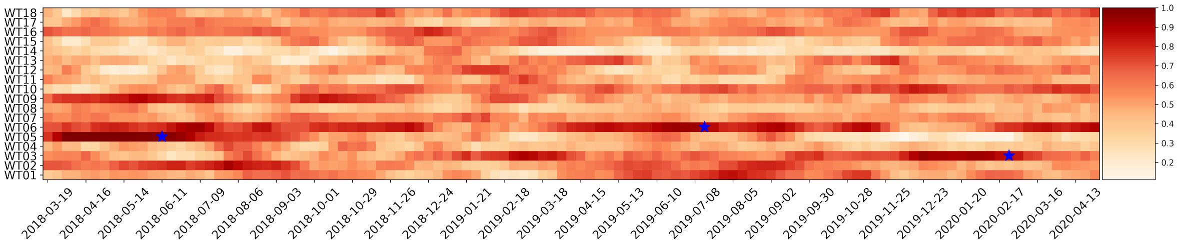


Figure 5. Main bearing health according to the proposed ensemble method. Faults are marked with a blue star.

To make a decision to trigger or not an alarm, a decision threshold is needed. Values of the indicator above the threshold will activate an alarm, and values below the threshold will be considered normal. To better visualize the results, a saturation mask is proposed in the following manner: all indicator values below the threshold are saturated to 0, and all indicator values above or equal to the threshold are saturated to 1. Figure 6 shows the ensemble indicator after applying the saturation mask with a decision threshold value of 0.85. It can be observed that the three main bearing faults in the park (highlighted with the blue stars) are warned at least 5 months in advance. On the other hand, there are false alarms at WT01, WT02, and WT09.

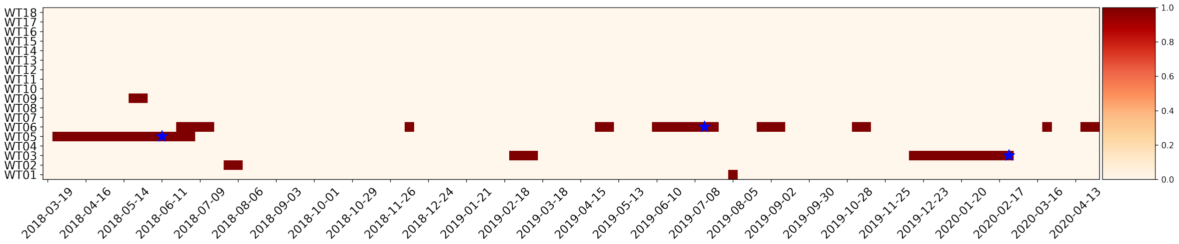


Figure 6. Saturation mask, with a decision threshold value of 0.85, used to determine whether a turbine should be reviewed or not. Faults are marked with a blue star.

The value of the decision threshold (DT) has a great impact on the final results. Figure 7 shows the DT value with respect to the different metrics. It is observed that, in general, higher values of the DT lead to better performance metrics, and it is noteworthy that for the same value of the DT the best performance metrics are clearly obtained with the ensemble method (except for low values of the threshold that have no practical relevance as they would lead to an excessive number of false alarms).

To further analyze the impact of the DT on the final results, Table 6 shows the confusion matrix information with respect to the DT value for the ensemble method. It is clear that low values of the threshold have no practical application since they lead to diagnose always as faulty the WT, since all results are positive (either TP or FP). Increasing the value of the DT diminishes the number of FP (false alarms), and consequently increases the number of TN (correctly classified healthy WTs). When the DT is equal to 0.95 only 28 instances are wrongly classified as FN (that is, faulty instances wrongly classified as healthy). When the DT is equal to 0.90 there are 58 instances wrongly classified as FP (false alarm), however, all faulty instances are correctly detected. In this case, when DT = 0.9, all faults are detected at the expense of having some false alarms. The wind park operator should compare the cost of unnecessary checks (due to false alarms) with respect to the savings of early warning of the main bearing fault to decide the best DT to be used.

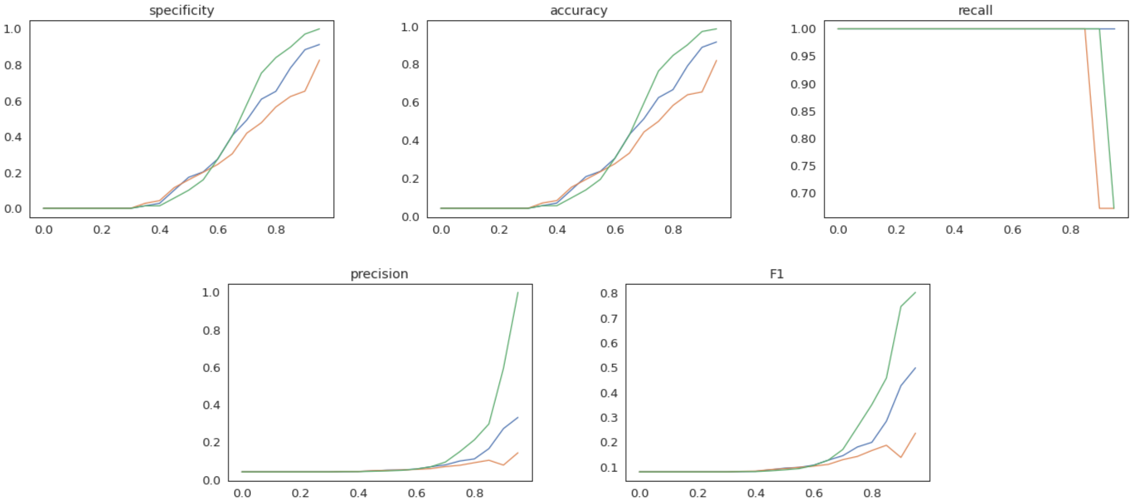


Figure 7. Decision threshold (DT) value vs. the different model indicators for the specificity (**top left**), accuracy (**top middle**), recall (**top right**), precision (**bottom left**), and F1 score (**bottom right**) metrics. The blue line corresponds to the anomaly indicator, the orange line to the normality indicator, and the green line to the ensemble indicator.

Table 6. Confusion matrix of the ensemble method results with respect to the decision threshold (DT).

DT	TP	FP	FN	TN
0.00	85	1967	0	0
0.05	85	1967	0	0
0.10	85	1967	0	0
0.15	85	1967	0	0
0.20	85	1967	0	0
0.25	85	1967	0	0
0.30	85	1967	0	0
0.35	85	1939	0	28
0.40	85	1939	0	28
0.45	85	1854	0	113
0.50	85	1769	0	198
0.55	85	1654	0	313
0.60	85	1427	0	540
0.65	85	1173	0	794
0.70	85	829	0	1138
0.75	85	485	0	1482
0.80	85	315	0	1652
0.85	85	201	0	1766
0.90	85	58	0	1909
0.95	57	0	28	1967

Finally, Figure 8 and Table 7 compare the performance metrics among the different single models and the ensemble model with respect to different DT values. In particular, Figure 8 shows a bar plot of the used metrics for three different specific values of DT, namely 0.6, 0.85, and 0.95. It is clear that the ensemble strategy outperforms in all cases the single methods, being outstanding the result in precision for DT=0.95 where a value of 1.0 is achieved with the ensemble but values lower than 0.5 are obtained with the normality and anomaly models by themselves. Table 7 gives a further detailed comparison of the performance metrics for values of the DT from 0 to 0.95 with increments of 0.05. Note that for values of the DT equal or bigger than 0.7 the best performance metrics are obtained with the ensemble method, clearly showing the advantage of this method with respect to the single models.

Table 7. Metrics comparison among the single and ensemble methods with respect to the decision threshold (DT).

	Recall			Specificity			Accuracy			Precision			F1		
DT	***Isolation Forest***	***Neural Network***	***Ensemble***	***Isolation Forest***	***Neural Network***	***Ensemble***	***Isolation Forest***	***Neural Network***	***Ensemble***	***Isolation Forest***	***Neural Network***	***Ensemble***	***Isolation Forest***	***Neural Network***	***Ensemble***
0.00	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.05	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.10	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.15	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.20	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.25	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.30	1.0	1.000	1.000	0.000	0.000	0.000	0.041	0.041	0.041	0.041	0.041	0.041	0.080	0.080	0.080
0.35	1.0	1.000	1.000	0.014	0.028	0.014	0.055	0.069	0.055	0.042	0.043	0.042	0.081	0.082	0.081
0.40	1.0	1.000	1.000	0.028	0.043	0.014	0.069	0.082	0.055	0.043	0.043	0.042	0.082	0.083	0.081
0.45	1.0	1.000	1.000	0.101	0.115	0.057	0.138	0.152	0.096	0.046	0.047	0.044	0.088	0.089	0.084
0.50	1.0	1.000	1.000	0.173	0.159	0.101	0.208	0.193	0.138	0.050	0.049	0.046	0.095	0.093	0.088
0.55	1.0	1.000	1.000	0.203	0.201	0.159	0.236	0.234	0.194	0.051	0.051	0.049	0.098	0.098	0.093
0.60	1.0	1.000	1.000	0.275	0.245	0.275	0.305	0.276	0.305	0.056	0.054	0.056	0.107	0.103	0.106
0.65	1.0	1.000	1.000	0.406	0.304	0.404	0.430	0.332	0.428	0.068	0.058	0.068	0.127	0.110	0.127
0.70	1.0	1.000	1.000	0.492	0.419	0.579	0.513	0.443	0.596	0.078	0.069	0.093	0.145	0.129	0.170
0.75	1.0	1.000	1.000	0.608	0.477	0.753	0.624	0.499	0.764	0.099	0.076	0.149	0.180	0.142	0.260
0.80	1.0	1.000	1.000	0.652	0.565	0.840	0.666	0.583	0.846	0.110	0.090	0.212	0.199	0.166	0.351
0.85	1.0	1.000	1.000	0.782	0.623	0.898	0.791	0.639	0.902	0.165	0.103	0.297	0.284	0.187	0.458
0.90	1.0	0.671	1.000	0.884	0.653	0.971	0.889	0.654	0.972	0.272	0.077	0.594	0.427	0.138	0.746
0.95	1.0	0.671	0.671	0.913	0.826	1.000	0.917	0.820	0.986	0.332	0.143	1.000	0.499	0.236	0.803

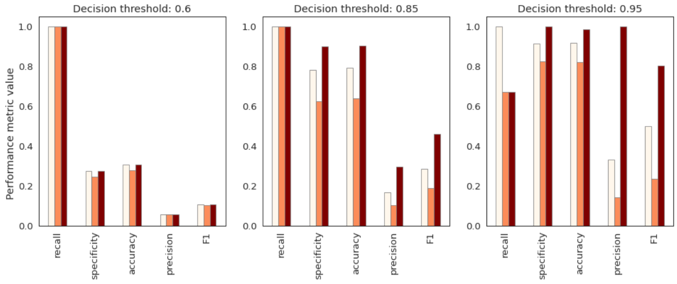


Figure 8. Performance metrics comparison for three different decision threshold values. The white bar represents the anomaly detection indicator, the orange one is for the normality indicator, and the red one for the ensemble indicator.

5. Conclusions

In this work an ensemble method for main bearing fault diagnosis has been proposed, implemented, and validated on a real under-production wind park composed by 18 WTs. The ensemble combines a normality model at WT level (i.e., a specific model is trained for each WT in the park) with an anomaly detection model at wind park level by using only SCADA data and past work order logs.

The stated ensemble method only requires healthy data to be trained (since it is not a supervised approach), and thus the methodology can be applied to any WT, even when there are no past records about the fault under study. The obtained results show that for the same value of the DT the best performance metrics are clearly obtained with the ensemble method in comparison to the single methods. Furthermore, the proposed strategy is able to warn at least 5 months in advance (giving enough time to the wind park operator to organize logistics and minimize downtime) the three main bearing faults present in the wind park during the test set, with few false alarms.

The DT has a great influence on the final results, thus it is strongly advised that the wind park operator analyzes the cost of unnecessary checks (due to false alarms) with respect to the savings of early warning of the main bearing fault to decide the best DT to be employed.

Two future work directions are envisioned. First, the extended isolation forest (EIF) presented in [30] will be incorporated to investigate the possible improvement in the results. Second, gearbox faults will be studied, as gearboxes tend to fail prematurely in WTs, and their replacement is very costly (the outage can last between a few days to months, depending on crane and parts availability). The wind park SCADA data and work order logs used in this research also contain gearbox faults, thus future work will address this type of fault, taking as starting point the ensemble method proposed in this work.

Author Contributions: All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially funded by the Spanish Agencia Estatal de Investigación (AEI) - Ministerio de Economía, Industria y Competitividad (MINECO), and the Fondo Europeo de Desarrollo Regional (FEDER) through the research project DPI2017-82930-C2-1-R; by the Generalitat de Catalunya through the research project 2017 SGR 388; by Centro para el Desarollo Tecnológico Industrial, grant number CDTI-IDI 20191294; and by Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), grant number DOCTORADO AGAUR-2017-DI 004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data required to reproduce these findings cannot be shared at this time as it is proprietary.

Acknowledgments: The authors express their gratitude and appreciation to the Smartive company (http://smartive.eu/, accessed on 5 July 2021), as this work would not have been possible without their support and ceded wind park data.

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

References

1. Reader, G.T. Energy, Renewables Alone? In *Sustaining Resources for Tomorrow*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–45.
2. Europe, W. *Wind Energy in Europe in 2018—Trends and Statistics*; Wind Europe: Brussels, Belgium, 2019.
3. Agency, I.R.E. *Renewable Power Generation Costs in 2019*; RENA: Abu Dhabi, United Arab Emirates, 2020.
4. Stehly, T.J.; Beiter, P.C. *2018 Cost of Wind Energy Review*; Technical Report; National Renewable Energy Lab. (NREL): Golden, CO, USA, 2020.

5. Kumar, A.; Shankar, R.; Thakur, L.S. A big data driven sustainable manufacturing framework for condition-based maintenance prediction. *J. Comput. Sci.* **2018**, *27*, 428–439. [CrossRef]
6. Ziegler, L.; Gonzalez, E.; Rubert, T.; Smolka, U.; Melero, J.J. Lifetime extension of onshore wind turbines: A review covering Germany, Spain, Denmark, and the UK. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1261–1271. [CrossRef]
7. Leahy, K.; Gallagher, C.; O'Donovan, P.; O'Sullivan, D.T. Issues with data quality for wind turbine condition monitoring and reliability analyses. *Energies* **2019**, *12*, 201. [CrossRef]
8. Vidal, Y.; Pozo, F.; Tutivén, C. Wind turbine multi-fault detection and classification based on SCADA data. *Energies* **2018**, *11*, 3018. [CrossRef]
9. Leahy, K.; Hu, R.L.; Konstantakopoulos, I.C.; Spanos, C.J.; Agogino, A.M.; O'Sullivan, D.T. Diagnosing and predicting wind turbine faults from SCADA data using support vector machines. *Int. J. Progn. Health Manag.* **2018**, *9*, 1–11.
10. Zhao, Y.; Li, D.; Dong, A.; Kang, D.; Lv, Q.; Shang, L. Fault prediction and diagnosis of wind turbine generators using SCADA data. *Energies* **2017**, *10*, 1210. [CrossRef]
11. Helbing, G.; Ritter, M. Deep Learning for fault detection in wind turbines. *Renew. Sustain. Energy Rev.* **2018**, *98*, 189–198. [CrossRef]
12. Cambron, P.; Masson, C.; Tahan, A.; Pelletier, F. Control chart monitoring of wind turbine generators using the statistical inertia of a wind farm average. *Renew. Energy* **2018**, *116*, 88–98. [CrossRef]
13. Li, Y.; Wu, Z. A condition monitoring approach of multi-turbine based on VAR model at farm level. *Renew. Energy* **2020**, *166*, 66–80. [CrossRef]
14. Astolfi, D.; Castellani, F.; Lombardi, A.; Terzi, L. Multivariate SCADA Data Analysis Methods for Real-World Wind Turbine Power Curve Monitoring. *Energies* **2021**, *14*, 1105. [CrossRef]
15. Hart, E.; Turnbull, A.; Feuchtwang, J.; McMillan, D.; Golysheva, E.; Elliott, R. Wind turbine main-bearing loading and wind field characteristics. *Wind. Energy* **2019**, *22*, 1534–1547. [CrossRef]
16. Hart, E.; Clarke, B.; Nicholas, G.; Kazemi Amiri, A.; Stirling, J.; Carroll, J.; Dwyer-Joyce, R.; McDonald, A.; Long, H. A review of wind turbine main bearings: Design, operation, modelling, damage mechanisms and fault detection. *Wind. Energy Sci.* **2020**, *5*, 105–124. [CrossRef]
17. Beretta, M.; Julian, A.; Sepulveda, J.; Cusidó, J.; Porro, O. An Ensemble Learning Solution for Predictive Maintenance of Wind Turbines Main Bearing. *Sensors* **2021**, *21*, 1512. [CrossRef] [PubMed]
18. Encalada-Dávila, Á.; Puruncajas, B.; Tutivén, C.; Vidal, Y. Wind Turbine Main Bearing Fault Prognosis Based Solely on SCADA Data. *Sensors* **2021**, *21*, 2228. [CrossRef]
19. McKinnon, C.; Carroll, J.; McDonald, A.; Koukoura, S.; Infield, D.; Soraghan, C. Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data. *Energies* **2020**, *13*, 5152. [CrossRef]
20. Zhang, Z. Missing data imputation: Focusing on single imputation. *Ann. Transl. Med.* **2016**, *4*, 9. [CrossRef]
21. Kayri, M. Predictive abilities of bayesian regularization and Levenberg–Marquardt algorithms in artificial neural networks: A comparative empirical study on social data. *Math. Comput. Appl.* **2016**, *21*, 20. [CrossRef]
22. Foresee, F.D.; Hagan, M.T. Gauss-Newton approximation to Bayesian learning. In Proceedings of the International Conference on Neural Networks (ICNN'97), IEEE, Houston, TX, USA, 9–12 June 1997; Volume 3, pp. 1930–1935
23. Bartilson, D.T.; Jang, J.; Smyth, A.W. Finite element model updating using objective-consistent sensitivity-based parameter clustering and Bayesian regularization. *Mech. Syst. Signal Process.* **2019**, *114*, 328–345. [CrossRef]
24. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010; pp. 249–256
25. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422. [CrossRef]
26. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
27. Wang, G.; Jia, R.; Liu, J.; Zhang, H. A hybrid wind power forecasting approach based on Bayesian model averaging and ensemble learning. *Renew. Energy* **2020**, *145*, 2426–2434. [CrossRef]
28. Lee, J.; Wang, W.; Harrou, F.; Sun, Y. Wind Power Prediction Using Ensemble Learning-Based Models. *IEEE Access* **2020**, *8*, 61517–61527. [CrossRef]
29. Wu, Z.; Lin, W.; Ji, Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access* **2018**, *6*, 8394–8402. [CrossRef]
30. Hariri, S.; Kind, M.C.; Brunner, R.J. Extended Isolation Forest. *arXiv* **2018**, arXiv:1811.02141.

5. Information Loss and SCADA Limitations

5.1 Limitations of SCADA data

SCADA utilization has greatly grown in the research and industrial field thanks to its wide availability and low cost. Nonetheless, concerns on the low frequency of SCADA —which can hinder the capability of modelling turbine conditions— have been raised by various researchers [39], [95]–[97].

Overall, the most evident issue with SCADA its the low time resolution of the data —that is commonly of one record every 5–10 minutes. Moreover, not all systems are properly equipped with dedicated sensors needed to train predictive algorithms. For example, it is common to have temperature signals of the drive-train components, whereas information regarding the torque, acceleration and displacements of bearings are uncommon. Pitch and yaw systems fare much worse, as often only the bare minimum set of control parameters —operating angles, but no pressure nor temperatures— are recorded making it very difficult to infer their conditions. While previous works have shown that results from high–frequency SCADA are more precise and the resulting models more accurate, no attempts have been made to quantify and determine the amount of information that is lost due to data aggregation.

As detailed in Ref. [11] SCADA data is recorded at a high frequency by a vast network of sensors, then it is aggregated to lower frequencies both at the turbine and windfarm level, and finally transmitted to a control center. The aggregation process allows to transfer data easily —due to reduced volume— and minimize the memory footprint of the databases where information is stored. While sensible from an operative perspective, this is not desirable from a modeling standpoint as important bits of information are inevitably lost.

5.2 Quantification of the Information Loss Resulting from Temporal Aggregation of Wind Turbine Operating Data

The proposed article deals with the limitations of SCADA data, more specifically with distortions introduced by the aggregating process that transform high–frequency SCADA data to the regular one in use in the industry. Specifically, this research aims to provide a quantitative framework to address the loss of information during aggregation of data. The objective is to better understand information loss and improve data storage policies.

High–frequency SCADA —i.e. 1 Hz data— is analyzed simulating the aggregation process and tracking differences between the original and rescaled signal. Methods based on statistics and an *"ad–hoc"* framework are compared to determine the signals that are more affected by the aggregation. Moreover, the effect of external conditions, i.e. the windspeed is evaluated. Finally, a graph showing the dependency between information loss and aggregation period is presented.

The results showed that the typical aggregation frequency of SCADA data results in a significant loss of information for signals describing electrical measurements (i.e. grid frequency, voltages) and wind. On the other hand, main bearing, ambient temperature or yaw angles retain most of their information regardless of the signal frequency. Another important insight from the paper is that for most signals information loss does not follow a linear decay and aggregation periods around 100–200 s allow to reduce the volume of data without scarifying excessively information content.

Contributions

The key contribution of this research is in **providing a framework** to use when analyzing the effect of signal aggregation and the resulting loss of information. This was an unanswered topic in the wind–energy literature. This framework has been used to answer three relevant question:

1. *How much information is lost with reduced temporal resolution?*
2. *Do external conditions have an effect on information loss?*
3. *What is the recommended aggregation frequency?*

The presented framework and the conclusions of this study aim to **raise awareness** and **provide a methodology** to assess limitations of SCADA data. Considering the growing importance of SCADA in predictive maintenance of wind turbines it could be useful to reconsider storage policies favoring the quality of the signal.

applied sciences

MDPI

Article

Quantification of the Information Loss Resulting from Temporal Aggregation of Wind Turbine Operating Data

Mattia Beretta [1,2,3,*], Karoline Pelka [2,*], Jordi Cusidó [3,4,*] and Timo Lichtenstein [2,*]

[1] Unitat Transversal de Gestió de l'Àmbit de Camins (UTGAC), Universitat Politécnica de Catalunya (UPC), 08034 Barcelona, Spain
[2] Fraunhofer Institute for Wind Energy Systems (Fraunhofer IWES), 30159 Hannover, Germany
[3] SMARTIVE S.L., 08204 Sabadell, Spain
[4] Enginyeria de Projectes i de la Construcció EPC, Universitat Politécnica de Catalunya, 08028 Barcelona, Spain
[*] Correspondence: mattia.beretta@upc.edu (M.B.); karoline.pelka@iwes.fraunhofer.de (K.P.); jordi.cusido@upc.edu (J.C.); timo.lichtenstein@iwes.fraunhofer.de (T.L.)

Abstract: SCADA operating data are more and more used across the wind energy domain, both as a basis for power output prediction and turbine health status monitoring. Current industry practice to work with this data is by aggregating the signals at coarse resolution of typically 10-min averages, in order to reduce data transmission and storage costs. However, aggregation, i.e., downsampling, induces an inevitable loss of information and is one of the main causes of skepticism towards the use of SCADA operating data to model complex systems such as wind turbines. This research aims to quantify the amount of information that is lost due to this downsampling of SCADA operating data and characterize it with respect to the external factors that might influence it. The issue of information loss is framed by three key questions addressing effects on the local and global scale as well as the influence of external conditions. Moreover, recommendations both for wind farm operators and researchers are provided with the aim to improve the information content. We present a methodology to determine the ideal signal resolution that minimized storage footprint, while guaranteeing high quality of the signal. Data related to the wind, electrical signals, and temperatures of the gearbox resulted as the critical signals that are largely affected by an information loss upon aggregation and turned out to be best recorded and stored at high resolutions. All analyses were carried out using more than one year of 1 Hz SCADA data of onshore wind farm counting 12 turbines located in the UK.

Keywords: SCADA; wind energy; operating data; high frequency; information loss; data storage; downsampling; temporal aggregation

check for updates

Citation: Beretta, M.; Pelka, K.; Cusido, J.; Lichtenstein, T. Quantification of the Information Loss Resulting from Temporal Aggregation of Wind Turbine Operating Data. *Appl. Sci.* **2021**, *11*, 8065. https://doi.org/10.3390/app11178065

Academic Editor: Daniel Villanueva Torres

Received: 5 August 2021
Accepted: 27 August 2021
Published: 31 August 2021

1. Introduction

In modern wind turbines, a plethora of operating data are acquired with high temporal frequency [1,2] by a vast number of sensors [3,4]. However, usually only a selection of these sensor data is stored. Furthermore, the data are typically aggregated as 10-min average values, sometimes accompanied by the standard deviations or the maxima and minima measured in these intervals. This temporal aggregation of a signal, also referred to as downsampling, saves a lot of space upon storage and reduces the bandwidth needed when transferring the data, both connected to cost savings. Unfortunately, much of the information on short timescales that might be valuable to better model and track the behavior and condition of wind turbines is inevitably lost in this process [4–6].

Aggregating data induces an information loss regardless of the source of data, even though its impact depends on the downsampling rate with respect to the behavior and the resolution of the raw signal. Albeit, the consequences arising from this signal conversion depend strongly on the further use of the data. Understanding these consequences, by knowing the properties of a signal after its transition to lower resolutions, will therefore help to optimize both data storage and costs while at the same time providing the best possible signal quality for analytic investigations.

In this study we investigate operating data of wind turbines recorded by the supervisory control and data acquisition system (SCADA). SCADA is a control system which among other functionalities allows for monitoring of wind turbines and receives input from a net of sensors that measure various operating variables such as wind speed, active power, temperatures, pressures, speeds, and environmental conditions through time. As we only investigate operating data, we will refer to "SCADA operating data" as simply "SCADA data" throughout the paper. Please note that although we developed the methods and algorithms to quantify information loss for wind turbine data, they are not necessarily restricted to our use case. While the considerations regarding specific consequences are limited to the wind energy field only, the approach can also be extended to other fields of application.

In many technological sectors data are considered a key asset to foster growth and innovation—wind energy is no exception. In the need of clean energy, wind power prospers globally: During the year 2020, 111 GW of new capacity were installed worldwide [7]. In Europe alone, wind power capacity amounted to 220 GW by the end of 2020. It is desired to grow between 80 and 105 GW over 2021–2025 out of which 29 GW are planned to be installed offshore [8]. Therefore, using data for better and more efficient operational strategies will be pivotal to further reduce the costs and to sustain the competitiveness of wind production compared to conventional energy sources. On the one hand, a profound data basis can support methods for accurate power output predictions. An overview about possible wind energy forecasts was given by Okumus and Dinler [9]. On the other hand, monitoring the health status of turbines supports the improvement of reliability by understanding and anticipating failures. The current state of using SCADA data for condition monitoring was summarized by Tautz-Weinert and Watson [4]. Each prevented turbine fault will avert a subsequent standstill and an involved loss of revenue. In the end this also reduces the costs for operation and maintenance (O&M) that currently account for up to 30% of an onshore wind turbine's levelized cost of electricity (LCOE) and up to 25% of the much higher LCOE of an offshore turbine [10]. These costs are thus still a major burden for the wind energy industry.

Research can help to minimize such costs by approaches of early fault detection or health monitoring methods. One possible solution addresses mechanical components by monitoring, e.g., the drivetrain or transmission elements through rich high-frequency data. Typically the sensors used are not part of the standard equipment of a turbine but of a dedicated condition monitoring system (CMS). Therefore, such a system needs to be installed explicitly and this is associated with further costs [11]. Occasionally, such high-resolution operating data are—partly—available from SCADA systems and could be used for the same purpose.

As stated earlier, data storage and transfer are associated with expenses. Therefore, for a cost optimization it is necessary to find a trade-off between reducing the amount of data, i.e., signal scope or resolution, and retaining enough information content to support O&M strategies. To illustrate the information loss Figure 1 shows the outcome of temporal aggregations of SCADA data at different time resolutions. Large aggregation intervals in the order of 300 to 600 s are not able to capture all nuances of the signal as local minima and maxima are flattened in the mean value curves. Clearly, this is only an example obtained for the wind speed and one temperature of the transmission shaft bearings, but similar behaviors can be observed for all signals.

Nevertheless, the figure gives an idea of the possible effects of temporal aggregation. A thorough study of the phenomena of information loss due to temporal data aggregation in the context of wind turbine SCADA data will be presented throughout this paper. The objective is to provide a deep analysis of the crucial points when aggregating SCADA data and to quantify the span of the information loss phenomenon. We also want to support turbine operators with a framework that allows them to take better decisions in terms of

SCADA data storage and aggregation policies. We break the general motivation down into these three specific research questions:

- Q1: How much information is lost with reduced temporal resolution?
- Q2: Do external conditions have an effect on information loss?
- Q3: What is the recommended aggregation frequency?

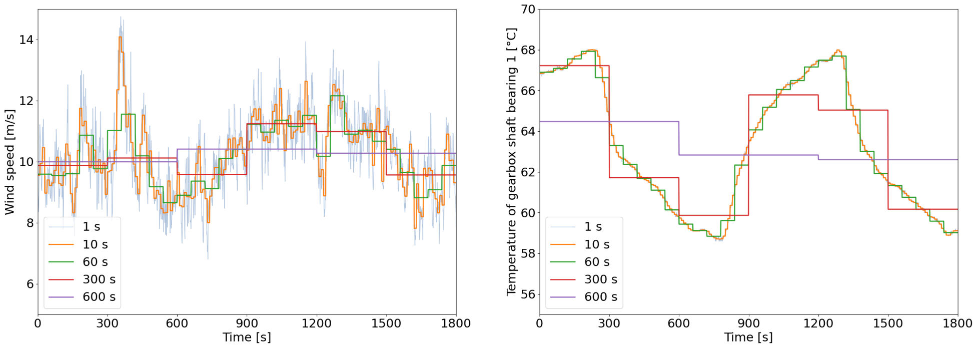


Figure 1. Illustration of temporal aggregation of SCADA operating data at various resolutions: On the left, a highly dynamic signal, the wind speed, loses much of its information even at low resolutions, while the original data of a slowly changing signal on the right, a temperature of the gearbox shaft bearing, is hidden behind the curve of the aggregation resolution of 10 s.

This paper is structured as follows: Section 2 provides a review of related work dealing with wind turbine monitoring and the use of high-frequency data in wind and other application sectors. Section 3 details the data set used in this study. Then, Section 4 tackles the key questions of this paper. Each subsection guides through our analytical approaches to answer these question and the corresponding results obtained. A discussion of the salient points derived from this study is also included for each question. Finally, Section 5 draws conclusions and discusses limitations of the study, the most relevant information, and ideas for future work.

2. Previous Work

Studies in the wind energy sector using high-frequency SCADA data are still scarce as these data sets are rare. Industry practice and state of the art is using aggregated SCADA operating data as 10-min averages. Nonetheless, a few available publications show the potential of the utilization of high-frequency SCADA data [1,2,12–15].

Generally SCADA data can be used for a high variety of applications. Two examples for major tasks in the wind energy sector are the prediction of the power output and the assessment of the turbine or component health status. Both problems are crucial for wind power production as they allow wind farm owners to keep their turbines spinning more and reduce mismatches between promised and delivered energy.

Power production estimation is a very large field of study within wind energy. A wide array of methods are available to predict power output [16–18]. The approaches vary on the type of data that is fed to the algorithms as well as the strategy used to detect patterns in the data. SCADA data was used in Refs. [15,19]. Since there is typically at least both the wind speed and the generated power within the SCADA data set, power output could ideally be estimated directly. However, this proved to be a very challenging task. Physical models focus mostly on the accurate prediction of the wind speed [20,21]. By using the characteristic power curve of a turbine, it is then translated to a modelled power.

As various models base on SCADA data and the resolution of this data is usually 10 min, i.e., rather low, it is relevant to determine the limitations of such aggregated data to better understand the possible shortcomings when predicting the power output.

Gonzalez et al. discussed the advantages of using high-frequency SCADA data in conjunction with a quantile random forest as predictive maintenance tool based on power curve modeling. The SCADA data used had a resolution of 4 s and was utilized to compare the predicted performance with 10-min averaged signals. As the natural variability of turbine operation was better captured by the high-frequency data, it also resulted in improved predictions [13]. In a later study, the same authors conducted a sensitivity study on the performance of high-frequency SCADA data as performance monitoring tool. Various factors such as terrain complexity, seasonality, choice of input variable, and most relevantly the sampling rate of data were analyzed. An important conclusion was the observation that a higher resolution allows to create more reliable models and as a key take-away they proposed to determine how much of the dynamic behavior of the signal is lost due to averaging [2]. Furthermore, the frequency of SCADA data is important to correctly model wake effects in wind farms. An inaccurate evaluation of wakes leads to imprecision in estimating wind speed and subsequent turbulence, which ultimately results in a poor prediction of power output [22].

A second important aspect in turbine operation is predictive maintenance. Unexpected and sudden failures can be very expensive for turbine owners and, therefore, an assessment of the turbine or component status shows up beneficial. The available data, the monitored system, and the requirements greatly influence the design and choice of a predictive maintenance toolbox.

As previously mentioned the drivetrain and other mechanical components, such as bearings and shafts, can be monitored by measuring acceleration, displacement and vibration through specific sensors and via acoustics emissions by dedicated CMS [11]. Additionally, for electrical and electronic components it is possible to apply CMS by analyzing current signatures in search of anomalous patterns [23]. These approaches are all based on utilization of very rich data, characterized by high sampling rates in the order of kHz. Here, signal processing techniques such as Fast Fourier transform, Hilbert–Huang transform, or wavelets analysis can be adopted [24–27].

Nonetheless, alternatives to the implementation of dedicated CMS in a wind turbine exist. SCADA data, while available at a much lower frequency and far poorer in terms of information, has also proven as a valid and cheap instrument for turbine monitoring. A short investigation on the usability of high-frequency SCADA data for predictive maintenance has been carried out by Roberts et al. [14]. In order to obtain the condition of a turbine, a quite plain approach is to model the power output and compare it to its measured value. In this manner, defects can be detected—ybut not localized—when significant discrepancies are observed [28,29]. So-called normal behavior models are another popular mathematical approach. These are regression models designed to predict the value of a key variable, capable of capturing the status of the studied system. A set of input variables is fed to an algorithm and the difference between the predicted and measured value is tracked. Large deviations are marked as anomalies and can be inspected further [30]. This goal can, e.g., be pursued by physical models [31] or neural networks [32]. Alternative approaches based on anomaly detection and fusion of multiple indicators and alarm logs, addressing generator and main bearing failures, have been proposed in [33,34]. When located in a wind farm, also adjacent turbines can serve as a reference value [31].

Further research also investigated simulating load by means of SCADA data from a single turbine or even the farm [35,36], eventually serving as an input for residual useful lifetime (RUL) estimations. Using the data of a whole farm can also serve to reconstruct an optimized flow through the area of this farm [37].

Dealing with the loss of information when temporally aggregating data or reducing the sampling rate is not an issue that occurs exclusively in wind energy. Nowadays, in cars even more data is collected by their electronic control units. Processing these data faces a similar problem: Liu et al. investigated the effects of reducing the sampling rate of the driving data with a particular focus on so-called micro-driving decisions such as spontaneous accelerations [38]. One result of this highly dynamic behavior was that the amount of information does not decrease linearly when reducing the temporal resolution, but that there are resolution ranges where no further information is lost whereas it falls down rapidly for other ranges.

Having considered the state of the art regarding wind turbine monitoring and power prediction, this paper aims to advance the understanding of information loss and possible limitations of SCADA data. In particular, the effect of low temporal resolution is analyzed and a quantitative method to determine the amount of lost information is detailed. Using this framework it is possible to determine signals that are most affected by downsampling and identify the influence of seasonal behavior and differences between turbines. An analysis of the effect of wind speed on information loss is also discussed, determining which operating conditions are affected the most by information loss. Finally, a methodology to choose an optimal aggregation frequency for a given signal is presented allowing to minimize the data storage footprint, while retaining most of the relevant information. In comparison with other high-frequency investigations this research offers the advantage of a large dataset consisting of more than one year of 1 Hz operating data that supports our conclusions.

3. Data

In this research we consider operating data generated by the SCADA system of wind turbines. The data set is gathered from an onshore wind farm consisting of 12 turbines with a nominal power of 2 MW commissioned in 2017 and located in the UK. The investigated period covers 15 months of data collection, hence seasonality effects should be limited.

Various signals that measure and monitor operational and ambient conditions are available with a temporal resolution of 1 second. Signals that represent counters are excluded in this analysis. In total, 27 signals are evaluated including temperatures, pressures, speeds, voltages, currents, and pitch angles. Table 1 provides an overview of the investigated signals, partitioned into functional groups. Please note that in this study only operating data is considered.

Table 1. List of SCADA signals analyzed in this study.

Component Temperatures	Control Variables	Electrical Characteristics	Environmental Variables	Mechanical Characteristics
Generator bearing 1 Generator bearing 2 Generator stator Gearbox oil Gearbox shaft bearing 1 Gearbox shaft bearing 2 Main bearing Top box	Pitch angle blade 1 Pitch angle blade 2 Pitch angle blade 3 Yaw angle	Active power Current phase A Current phase B Current phase C Grid frequency Power factor Voltage phase A Voltage phase B Voltage phase C	Ambient temperature Nacelle temperature Wind direction Wind speed	Generator speed Rotor speed

The time stamps of the original dataset are not always exact to the tick of a second. Sensors, for various reasons, might record their values slightly early or late. For computational reasons we decided to adjust the timestamp downward to seconds with the typical floor-functions available. Therefore, in a few cases two different records were assigned to the same timestamp. In these cases only the earlier one of the two values is kept. Furthermore, for missing data no replacement, i.e., no imputation, was performed and generally no treatment for outliers was performed, except for the descriptive statistics in Section 4.1.1.

In order to examine the effect of temporal aggregation, throughout this paper downsampling of high-resolution data to a lower temporal resolution is accomplished by averaging if not stated otherwise. Table 2 depicts an example for the temporal aggregation of the wind speed.

Table 2. Illustration of the temporal aggregation scheme: The original data, i.e., t (1 s), is averaged over the lower temporal range and given a new timestamp at the beginning of this interval, see t (5 s) and t (10s).

t (1 s)	Wind Speed (m/s)	t (5 s)	Wind Speed (m/s)	t (10 s)	Wind Speed (m/s)
09:00:00	6.26	09:00:00	6.25	09:00:00	6.42
09:00:01	6.11				
09:00:02	6.17				
09:00:03	6.39				
09:00:04	6.32				
09:00:05	6.37	09:00:05	6.71		
09:00:06	6.66				
09:00:07	6.72				
09:00:08	6.72				
09:00:09	7.07				
09:00:10	7.29	09:00:10	...	09:00:10	...

4. Analyses

In this section, we present the applied methods addressing each research question Q1–Q3. Furthermore, the corresponding results are reported, described, and discussed.

4.1. Q1: How Much Information Is Lost with Reduced Temporal Resolution?

Understanding the effect of temporal aggregation, quantifying the induced information loss, and identifying the signals which are affected the most are important steps to find a trade-off between data volume and information content. Hence, storage policies can be defined and possibly additional space can be allocated for the signals that are not recommended for aggregation. The methods and approaches in this subsection address the guiding question, if data when sampled at a higher frequency contains richer information. In this study simple approaches based on statistics indicators and tests are replicated. Then, given their limitations a different method based on the analysis of the aggregation error is devised.

4.1.1. Comparison of Descriptive Statistics

The effect of different levels of aggregation of data is studied by comparing the values of a set of descriptive statistics. The objective is to identify global changes in the signals that are reflected in their range, central behavior and overall shape of distribution.

Methodology

In order to evaluate the effects of temporal aggregation, the first approach constituted the calculation of key descriptive statistics, which captured the central behavior, shape and dispersion of the data. To examine the effect of temporal aggregation, these statistics were computed for the non-aggregated, i.e., raw data (1 second of temporal resolution) and for temporally aggregated data with reduced time resolutions, namely 10 s, 60 s, 300 s, and 600 s. For each signal listed in Table 1 the following statistics were computed for all resolutions mentioned above: sample mean, median, maximum, minimum, standard deviation, first quartile, third quartile, skewness, and kurtosis. For the calculation specification of these quantifiers we refer to Ref. [39].

In contrast to the subsequent analyses, the calculations of this part were carried out on the entire dataset. To keep extreme outliers out of the scope for less distorted results, during this analysis the data for wind speed and generator speed were filtered for only positive values and temperatures for values <200 °C.

Results

Selected results are shown in Figure 2 for one turbine. The full list of computed statistics based on different temporal resolutions for the entire set of signals is provided in the

Supplementary Materials. Here, we present an appropriate graphical representation using a box-and-whisker-plot [40], which includes the mean (denoted by the green triangles), the median (denoted by a horizontal yellow bar in the box), the first quartile (denoted by the bottom edge of the box), the third quartile (denoted by the top edge of the box), the minimum (denoted by the bottom edge of the whisker) and the maximum (denoted by the top edge of the whisker). Moreover, the green bars in the plot represent the mean value ± the standard deviation. These are shown for four exemplary signals, namely wind speed, active power generation, generator speed, one temperature of the gearbox shaft bearings, and the voltage of phase A. Note that in this version of box-and-whisker-plots the whiskers represent the minimum and maximum of the underlying data and therefore include outliers.

To display the results of skewness and kurtosis we chose a simple scatter plot that is displayed in the second row of Figure 2 for the same signals mentioned above accordingly. As both measures are normalized to the standard deviation they are both plotted against the same dimensionless ordinate.

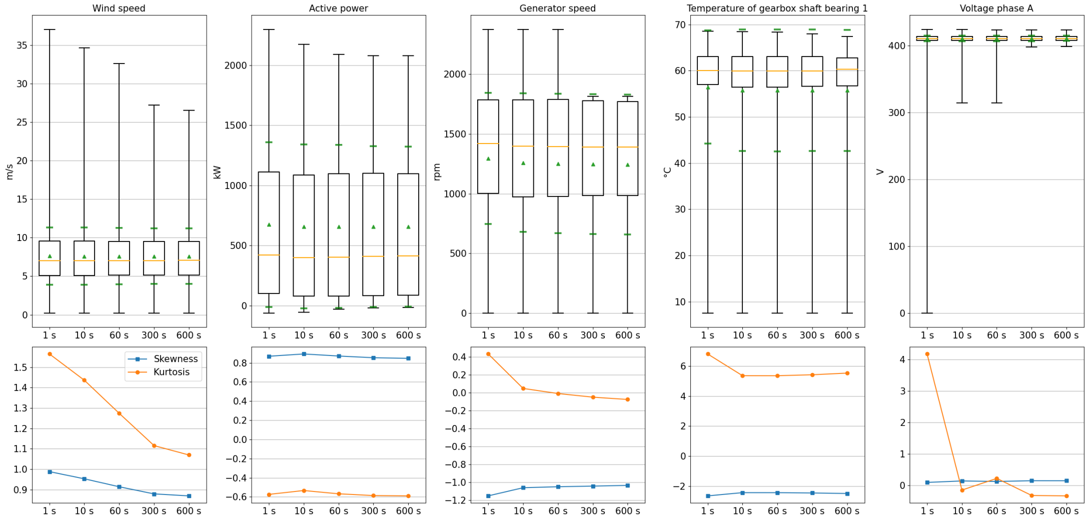


Figure 2. Illustration of the descriptive statistics of wind speed, active power generation, generator speed, temperature of gearbox shaft bearing 1, and voltage of phase A at different levels of temporal aggregation for wind turbine 5. On the top: box-and-whisker plots with the median depicted by the yellow bar, and the mean by the green triangles surrounded by ± the standard deviation as green bars. The black box displays the interquartile range, the black whiskers denote the maximum and minimum, including outliers. On the bottom: scatter plots of skewness and kurtosis normalized to standard deviation with interconnecting lines for better visibility.

From the representative statistics of the graphical presentation in Figure 2 and the additional data in the Supplementary Materials the following results can be obtained:

- Temporal aggregation had a pronounced effect on the maxima: with lower temporal resolution the maxima decreased. Especially for the wind speed we saw a distinct decreasing trend of the maxima (for this presented turbine and investigated time period a reduction of 10 m/s).
- A corresponding effect for the minima, i.e., an increasing trend, could not be seen for those variables with a defined lower boundary, e.g., 0 m/s for the wind speed. Here, lower boundaries were preserved. For other variables, a similar but much less pronounced effect was existent, e.g., for the active power generation. The one exception was the voltages, as their value only dropped from the reference value in the case of disturbances.

- Mean and median faintly declined for most signals with increasing temporal aggregation, though in the illustration no noticeable visual differences could be obtained.
- The standard deviation and the interquartile range (IQR) behaved similarly, although both also experienced only small changes: For some values they faintly decreased, e.g., for the wind speed and the active power, whereas for others like generator speed and the temperature of the gearbox shaft bearings they made a small increasing step between 1 s and 10 s.
- The values of skewness were rather close to each other for different levels of temporal aggregation. Although the base values scattered a lot, the median of their changes for all turbines and signals lay below 5% with respect to the raw signal. An explicit decline could only be seen for the wind speed.
- For the kurtosis a clear trend could only be observed for the wind speed and the voltages where there was a clear decline with increasing temporal aggregation. For the rest of the signals this was not as clear: for several signals there was almost no difference, some signal tended to increase for one turbine, but decreased for another. However, the median of all changes lay below 5% change in standard deviation with respect to its value at 1 s resolution.

In summary, the statistics of typically fast changing signals such as wind speed, active power, generator speed, and electrical signal were the most affected by downsampling as the values of their statistics varied widely with the resolution of the signal. Temperature signals, on the other hand, were far less affected by downsampling.

Discussion

Most prominent variations in the statistical analysis are the minima and maxima of the signals, displayed by the whiskers in Figure 2. Mostly, maxima tend to decrease with the length of the aggregation period, meaning that longer aggregations smooth out peak values of the signal possibly annihilating anomalous conditions. If peak values are strongly reduced from the data, it can be assumed that also short negative spikes will undergo the same behavior. Except for the voltages, this cannot be seen in most plots as the minima coincide with those periods in which turbines are not operating. Still, this reduction of peak values might negatively impact the possible performance of early-fault detection algorithms and more general models attempting to represent the behavior of turbines under peak load conditions. For these use cases, high frequency SCADA should be considered.

Mean and median values are almost constant with respect to the aggregation period length as the total weight of the values cannot be significantly shifted by averaging. The change of the standard deviation and the IQR behaves differently and can exhibit the following patterns: An increase means that several sudden outliers on short timescales are annihilated by the averaging process and, therefore, less data is distributed at the tails. A decrease is observed when the values move to the tails by the averaging process due to the majority of data inside an aggregation windows distributed at the tails.

The shape indicators skewness and kurtosis do not provide completely conclusive information. However, there seems to be a connection between the reduction of the maximum or minimum and a decrease of the kurtosis, as can be seen for the wind speed, the generator speed, and the voltage. If due to the aggregation process there are less remote outliers, the probability distribution of the data becomes less spiked. Consequently, we can observe a loss of short-time peaks in the data by the reduction of the kurtosis. In contrast, the interpretation of the skewness is not as simple. However, it should give a hint to the direction of the majority of outliers or peaks, respectively. For the wind speed the skewness becomes less negative, telling us, that most of the outliers in the positive direction for the 1 s data will be smoothed out with increasing aggregation time. For the generator speed of the exemplary turbine in Figure 2 more outliers from the negative direction are smoothed. Please note again, that these are not generally valid as, except for the wind speed, the behavior of the skewness of a signal varies from turbine to turbine in the given dataset.

Observing the behavior of a set of descriptive statistics provides a simple check on the effect of temporal aggregation. The proposed analysis attempts to capture various characteristics of the signal distribution, including its shape, dispersion, and central behavior. Overall, this observation of descriptive statistics does not answer the question on how much information is lost due to temporal aggregation. Changes in the range and standard deviation of the signals provide a general indication of the effects of temporal aggregation for a set of signals. Although pointing at the temporally critical signals, they fail to quantify precisely the loss of information. Moreover, this approach does not provide insights into the dynamic behavior of the signals as all indicators provide only a global perspective. Providing indications on the optimal frequency of a signal based on the value of descriptive statistics is particularly challenging.

4.1.2. Kolmogorov–Smirnov Test

Beside comparing descriptive statistics calculated for different levels of aggregation, inferential methods were applied in order to quantify a change of the distribution of the aggregated signal. Here, the Kolmogorov–Smirnov (KS) two sample test was used to ascertain whether the distribution of the temporally aggregated signal differed from the distribution of the non-aggregated signal.

Methodology

We only briefly introduce the idea and procedure of the test, for a detailed description please see [41].

In this work, the KS two sample test was applied and no assumption on the distribution of the underlying data was made. Suppose the data consisted of two independent samples, a first sample $X_1, X_2, ..., X_n$ of size n and a second sample $Y_1, Y_2, .., Y_m$ of size m. $F(x)$ and $G(x)$ denote their respective, unknown distribution functions. We wanted to test the hypothesis

$$H_0 : F(x) = G(x) \quad \text{vs.} \quad H_1 : F(x) \neq G(x). \tag{1}$$

Let $F_n(x)$ be the empirical distribution function based on the random sample $X_1, X_2, ..., X_n$ and let $G_m(x)$ be the empirical distribution function based on the random sample $Y_1, Y_2, .., Y_m$. Then, the test statistic D measures the maximum difference between the two empirical distribution functions and is defined as

$$D = \sup_x |F_n(x) - G_m(x)|. \tag{2}$$

From the test statistic D the p-value was derived, adjusting for the different sample sizes n and m. As a level of significance the typical value of $\alpha = 0.05$ was chosen.

If the p-value was smaller than the level of significance, i.e., $p < 0.05$, H_0 was rejected and we had sufficient evidence to say the aggregated data had another underlying distribution. If the p-value was greater than 0.05, then we did not have sufficient evidence to reject the null hypothesis and we could not draw any further conclusions. Please note that in our case there was an important difference of the test application when compared to its standard use. Here, we compared two samples that are known to come from the same generating process, as the tested signal was the same. Therefore, the test tried to determine whether the aggregation process changed the distribution of the signal such that the aggregated and original distribution no longer appeared similar.

Results

The test statistics D and the p-values were calculated for all signals based on incorporating the raw signal and several aggregated signals on different temporal resolutions of 10 s, 60 s, 300 s, and 600 s, equivalent to Section 4.1.1. The test was conducted with 10,000 h of randomly chosen samples of operating data. As the quantity of interest we present the derived p-values for an exemplary turbine, i.e., turbine 5, containing all investigated signals on the left side of Figure 3 and the minimum values of all turbines on the right. Data containing all numerical results are provided in the Supplementary Materials.

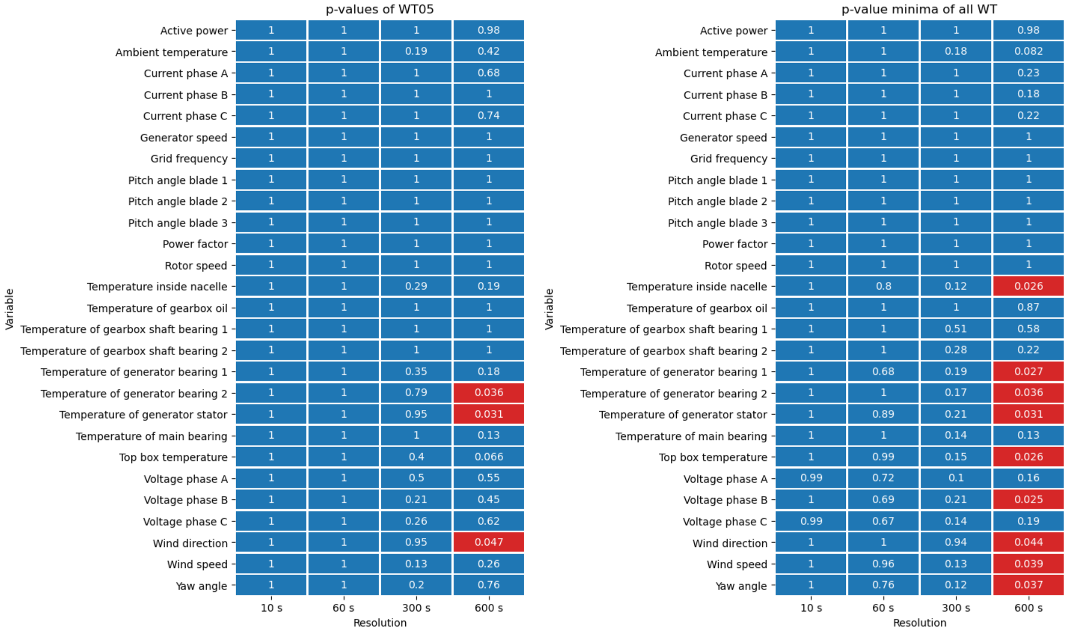


Figure 3. Calculated *p*-values of conducted KS tests for all signals at different levels of temporal aggregation for one exemplary turbine WT05 on the left and the minimum value of each cell of all turbines on the right. Cells with a red tint are below the level of significance, the blue ones above.

Regarding all turbines the majority of *p*-values for all signals were above the significance level of 0.05. Especially for higher resolutions <60 s the *p*-values even tended to keep around 1, making it very unlikely that both the raw and aggregated value came from two different underlying distributions. There was even a set of signals whose *p*-values were always 1. These were the active power generation, grid frequency, generator speed, all pitch angles, power factor, and rotor speed. The turbine on the left of the figure had explicitly been chosen because it featured *p*-values below 0.05 for certain channels for aggregated data of 600 s resolution. As displayed on the right of Figure 3, regarding the whole wind farm for the following signals the KS test showed *p*-values below 0.05 at least for one turbine: the temperature inside nacelle (1×), both generator bearing temperatures as well as the generator stator temperature (each 1×), the top box temperature (1×), voltage phase B (1×), wind direction (2×), wind speed (2×), and yaw angle (1×). Thus, for the last mentioned group of signals, resampling to 600 s could alter the data in a way such that the underlying distribution of the data was no longer the same as for the raw data. One additional signal that also came close to our significance level with a *p*-value of 0.082 once within the scope of all turbines was the ambient temperature.

Discussion

From the results of all turbines, we extracted the minimum *p*-value for each signal and aggregation value in Figure 3. Of course, this is a rather unconventional approach with questionable statistical significance. Nevertheless, in this way the results of the KS test can be divided into three signal subgroups that are shown in Table 3: (1) signals for which the test yields *p*-values of 1 or only slightly below. (2) Signals with a resulting *p*-value that decreases with increasing aggregation time, but never falls below the chosen level of significance of 0.05. (3) Signals that fall below the level of significance at least once for all turbines.

Table 3. Signals sorted into bins of p_{min}. Here, p_{min} is the minimal p-value of a signal for all aggregation resolutions and turbines as already carried out in Figure 3.

$p_{\text{min}} \geq 0.98$	$0.98 > p_{\text{min}} \geq 0.05$	$p_{\text{min}} \leq 0.05$
Active power	Ambient temperature	Temperature inside nacelle
Generator speed	Current phase A	Temperature of generator bearing 1
Grid frequency	Current phase B	Temperature of generator bearing 2
Pitch angle blade 1	Current phase C	Temperature of generator stator
Pitch angle blade 2	Temperature of gearbox oil	Top box temperature
Pitch angle blade 3	Temperature of gearbox shaft bearing 1	Voltage phase B
Power factor	Temperature of gearbox shaft bearing 2	Wind direction
Rotor speed	Temperature of main bearing	Wind speed
	Voltage phase A	Yaw angle
	Voltage phase C	

For the first group of signals the p-value of the KS test was always much higher than our level of significance of 0.05, some even stayed permanently at 1. Although, from these results we can only conclude that the two samples, i.e., the raw data and the aggregated values, are not from two different underlying distributions, it is also a good sign, because it gives us an indication that most of the aggregated signals are not strongly disturbed with respect to the original signal for all resolutions. Other signals showed a decreasing p-value with increasing aggregation time. Still, most of the p-values were above our significance threshold. This could mean that, contrary to the first group, the signals deviate more and more with when decreasing the resolution. In the third group of signals, the KS test resulted in p-values below the level of significance at least once for all turbines. These low values of $p < 0.05$ occurred only for aggregation times of 600 s and only once or twice in the whole set of turbines. Therefore, they do not tell us that the respective signals are always altered heavily by the aggregation process. However, those data give us hints to signals that can show short term deviations in their time series that might be annihilated during mean value aggregation process. Therefore in return, these short term deviations only occur very rarely—in our case only for one or two turbines. Regarding the voltages, there was even only one prominent phase. The observed rarity also makes it hard to tell that the list of signals with eventual important short term deviations is complete. For example: All voltages should behave in the same manner.

As a conclusion, the KS test revealed deviations for signals of which some fall into a group of typically fast changing values, such as the wind speed and direction, and the measured voltages. The other group consists of some rather lagged signals with several temperatures as well as the yaw angle. These signals might contain valuable information on short timescales that is lost during aggregation while it might serve as an important input for future predictive methods such as early fault detection. Finding these short-time features will certainly be a task for the learning process of such methods. Here, the KS test might be helpful to identify the interesting signals and time ranges. However, the KS test falls short of giving quantifiable indications on which resolution to choose.

4.1.3. Local Error Approach

In this section we will address the information loss directly, conducting an analysis on the bare differences between the raw signal and the signals aggregated over different aggregation intervals.

Methodology

We defined the information loss as the difference between the original signal and its down-sampled signal, i.e., the local error, described by the loss function $\Lambda(t)_{\Delta t_{\text{agg}}}$. The calculation was carried out by further applying the absolute value on the difference, as described in Equation (3):

$$\Lambda(t)_{\Delta t_{\text{agg}}} = |s(t)_{\Delta t_{\text{agg}}} - s(t)_{\Delta t_{\text{org}}}| \tag{3}$$

In this equation Δt_{org} is the original or native resolution of the data before aggregation. The original signal values are defined as $s(t)_{\Delta t_{\text{org}}}$, the aggregated values as $s(t)_{\Delta t_{\text{agg}}}$ correspondingly. Note that $\Delta t_{\text{org}} = 1\,\text{s}$ for the present dataset. Within a sampling window all values $s(t)_{\Delta t_{\text{agg}}}$ were equal to the averaged raw signal of this window, also known as "value hold". Thus, the information loss $\Lambda(t)_{\Delta t_{\text{agg}}}$ always has the same resolution as the original data. A representation of the original and aggregated signal is provided in Figure 4.

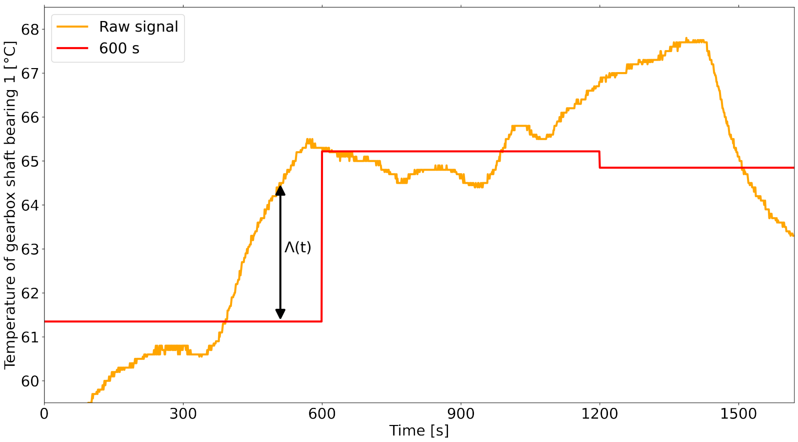


Figure 4. Exemplary illustration of the information loss $\Lambda(t)_{600\,\text{s}}$ of the gearbox bearing temperature for a temporal aggregation of 600 s and a raw signal in 1 s resolution.

Due to the diverse nature of physical quantities measured by the SCADA system, the local errors needed to be normalized to be compared across signals effectively. Normalization also facilitated the interpretation of the results, as it allowed to reason about information losses in relative terms, bypassing the necessity to know the typical operating range of a signal. In our analyses, the interquartile range (IQR) of the values of the raw time series was used as the normalization basis throughout the following parts of the paper if not declared otherwise. The calculated IQR values are listed in Table 4. We chose this range over other common approaches, e.g., the minimum to maximum distance, to deal with the presence of unavoidable outliers. At the same time the IQR normalization could deal with multi-modal signals that might alternate between two regime states, rarely taking in-between values. Here, working with the the standard deviation could lead to obscure results.

The analysis was performed on a set of 1000 hours of operation, i.e., approximately 0.8% of the entire data set, that was randomly sampled from the complete time series of all wind turbines. For this method the time resolutions varied from 5 to 600 s, the latter being the typical timescale of SCADA data for commercial wind farms.

Results

The resulting normalized information loss function values $\Lambda(t)_{\Delta t_{\text{agg}}}$ were grouped into bins of error size. Then, the percentage of data in each bin was calculated, providing us an overview about the severity of the information loss for each signal after an aggregation in various resolutions. In Figure 5 the results are presented: each signal was assigned one bar lined up on the x-axis, colors represent the aforementioned error bins. The margins of these bins are defined as fractions of the normalization basis. The height of each bar is given by the percentage of data in the corresponding error bin. The signals are sorted in ascending order from left to right on the horizontal axis according to the percentage of data in the lowest error bin of $[0, 0.025] \times$ IQR at the largest aggregation period of 600 s. Each signal is assigned a number in the bottom-most axes to allow for better orientation in the upper plots. This representation allowed us to yield an information loss added up from the local error, to compare different signals, and to determine the most affected signals. Moreover, it provided a quantification of the fraction of information that was lost for an assigned resolution for each signal. Tables with the numerical results are provided in the Supplementary Materials.

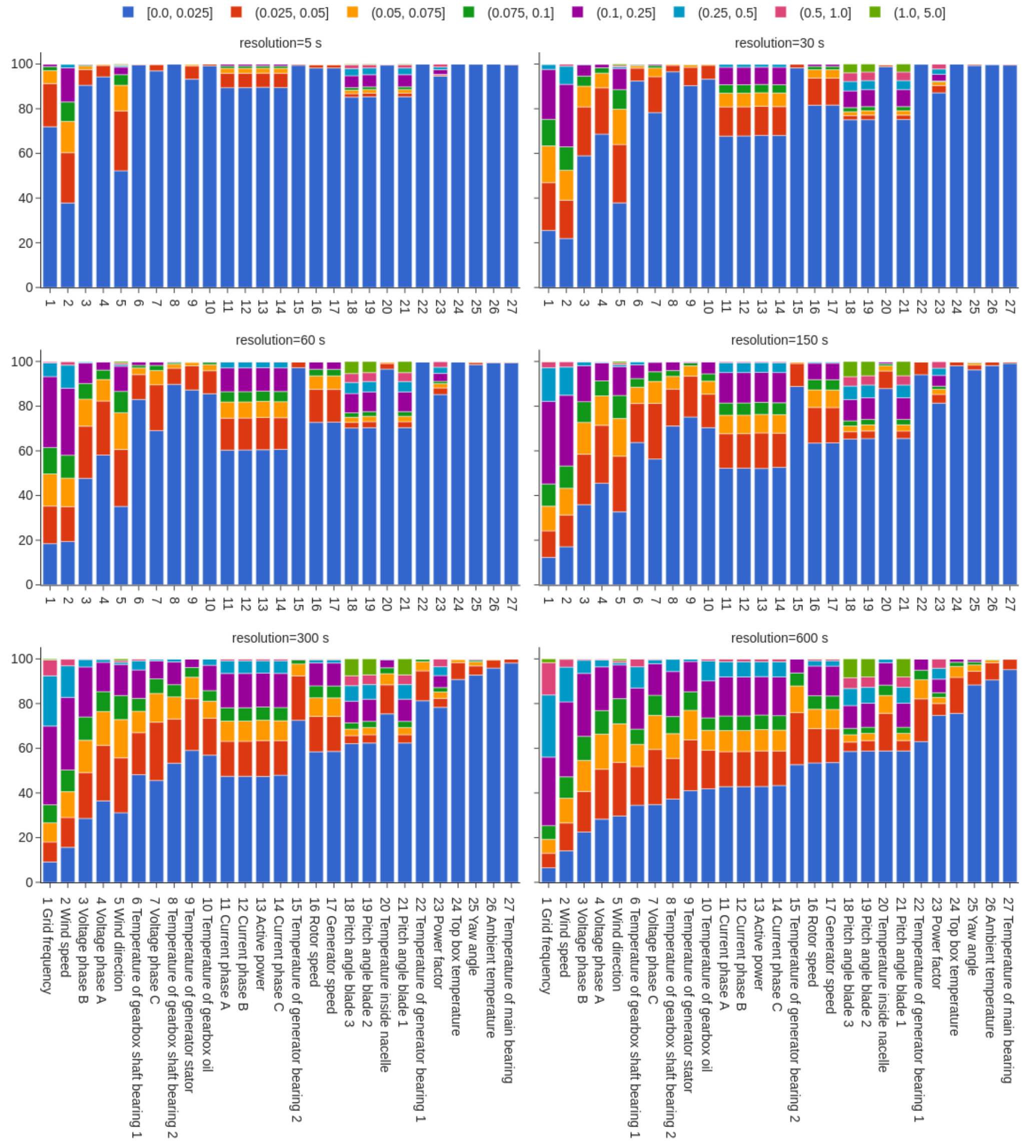


Figure 5. Information loss results for all the available signals and various aggregation ranges. Error values are normalized and discretized into bins, defined as fractions of the IQR value. Colors represent the error fraction bins. Signals are identified by a number in the bottom-most row for improving the readability in the upper panes.

Table 4. List of the available signals and corresponding interquartile range values used to normalize results for comparability. The IQR is defined as the difference between the third and first quartile of a distribution.

Signal Name	IQR	Unit
Active power generation	1052.1	kW
Ambient temperature	6.8	°C
Current phase A	843.0	A
Current phase B	842.0	A
Current phase C	838.2	A
Generator speed	804.5	rpm
Grid frequency	0.098	Hz
Pitch angle blade 1	1.6	deg
Pitch angle blade 2	1.6	deg
Pitch angle blade 3	1.5	deg
Power factor	1.998	-
Rotor speed	7.8	rpm
Temperature inside nacelle	12.0	°C
Temperature of gearbox oil	4.9	°C
Temperature of gearbox shaft bearing 1	6.0	°C
Temperature of gearbox shaft bearing 2	8.2	°C
Temperature of generator bearing 1	11.6	°C
Temperature of generator bearing 2	5.1	°C
Temperature of generator stator	6.0	°C
Temperature of main bearing	5.5	°C
Top box temperature	10.9	°C
Voltage phase A	5.5	V
Voltage phase B	6.4	V
Voltage phase C	6.1	V
Wind direction	151.5	deg
Wind speed	4.4	m/s
Yaw angle	153.4	deg

When inspecting the results displayed in Figure 5, it is possible to obtain behaviors of the signals, such as critical drops of the information content passing from one resolution to another. Additionally, signals that are heavily affected by information loss can be identified. Some key observations are highlighted as follows:

- The information loss severity, i.e., the maximum error occurring, varied greatly between the signals. Certain signals had roughly more than 1% of data with an error greater than 0.5 IQR , mainly environmental, electrical and control variables, such as frequency, currents, power factor, wind speed, and pitch angles. However, for the temperatures of gearbox shaft bearing 1 and the gearbox oil there was also a small amount of error above 0.5 IQR.
- Generally, temperatures were not particularly affected. Only a fraction of information was lost for the largest aggregation period. A noticeable exception was the aforementioned transmission signals, i.e., gearbox shaft bearing 1 and 2 as well as the gearbox oil temperature, which had less than 50% of the data included in the lowest error bin at 600 s aggregation resolution and, therefore, underwent a relatively high loss of information.
- Wind speed and electric signals underwent a drastic loss of information, even at the short aggregation periods (5 to 30 s). Wind speed data in particular featured only 40% of the data in the lowest error bin, i.e., an error $\leq$2.5% IQR.
- Excluding the wind measurements, the pitch angles, electrical characteristics, and generator and rotor speed, information loss was limited to $\approx$18% of data with an error of $>$0.025% IQR up to an aggregation interval of 60 s. Above this threshold of 60 s resolution, most signals began to lose a considerable amount of information, as the shrinking percentage of data in the lowest error bins indicated.

- Current and active power signals are strongly affected for resolutions above 5 s. Changing the aggregation from 5 s to 30 s causes a loss of >20% of the total data in the lowest error bin of $[0, 0.025] \times$ IQR. While the amount of data in this lowest error bin kept decreasing with a further reduction of the resolution, it was not as drastic as from 5 s to 30 s.
- The typical SCADA data resolution, i.e., 600 s, was not sufficient to correctly represent wind measurements, electrical signals, and the temperatures of gearbox and generator components as more than 20%, even >50% for the wind speed, of the data have losses greater than 0.1 IQR. On the other hand ambient temperature, main bearing, top box temperature, and yaw angle were barely affected retaining more than 80% of the data in error bins lower than 0.1 IQR
- Pitch angle values were occasionally affected by large differences between the aggregated and original signal. Approximately 10% of the data manifested losses superior to 1 IQR, even at short aggregation periods such as 30 s.
- The transition from 150 to 300 s caused a visible drop from approximately 90 to 70% of the size of the lowest error bin of the generator bearing temperature.

Behavior of Temperature Signals

Temperature signals constitute a special interest subgroup, as they are typically used as inputs for predictive maintenance to monitor the status of turbine components. Therefore, we conduct a separate investigation specifically for temperatures. Here, a normalization of results was not necessary as all temperatures already shared the same physical unit, i.e., degrees Celsius.

The results for selected aggregation times are presented in Figure 6. Except for the error bins now in °C, it is the same representation as in the previous part. All data are provided in the Supplementary Materials. The following observations are emphasized:

- Up to an aggregation period of 150 s information losses above 1 °C were almost nonexistent . More than 97% of the data were contained in the [0, 1 °C] error bin. Only gearbox shaft bearings and internal temperatures had a negligible amount of data in the second error bin.
- The typical SCADA resolution, i.e., 600 s, could be problematic for the temperatures of the gearbox shaft bearings, of the gearbox oil, and inside the nacelle, as only 80% of the data had an error below 1 °C.
- Gearbox shaft bearings and gearbox oil temperatures could occasionally exhibit information losses higher than 2 °C and more rarely higher than 3 °C for aggregation periods of 600 s.

Turbine-Dependency

The results of the previous calculation of $\Lambda(t)_{\Delta t_{agg}}$ could also be split between turbines, obtaining a breakdown of the information loss across the whole wind farm, as shown in Figure 7. This analysis allowed us to verify whether information loss was a condition imputable to the behavior of single turbines, or rather a generalized phenomenon affecting all turbines in the farm. Only a selection of temperature signals and resolutions is presented, the complete results table can be found in the Supplementary Materials. The following observations can be drawn:

- Not a single turbine or a set of turbines was responsible for the entire amount of information loss.
- There were variations in the amount of lost information across turbines. For example, observing the percentage of data of the gearbox shaft bearing temperature with an aggregation time of 600 s for which the aggregation error was below 1 ºC, the differences between turbines could vary within a range greater than 20%. In particular turbines WT04 and WT06 had slightly more than 60% of the data in the lowest error bin versus turbine WT02, WT03, and WT11 that had approximately 90% of the data within the 1 ºC error range.

- The differences between turbines were principally only visible for transmission related signals, i.e., gearbox oil and gearbox shaft bearing temperatures. For all other temperatures differences between turbines were not noticeable, as information losses were overall very limited.
- Aggregating the signal at a lower, yet still coarse time resolution, i.e., 300 s, reduced the differences between turbines. The variation range was closer to 10% in this case.
- While shorter aggregation periods reduced the differences between turbines, it did not change the relative impact of information loss within the wind farm. This means that the most affected turbines at 300 s aggregation were also the ones showing greater losses at 600 s.

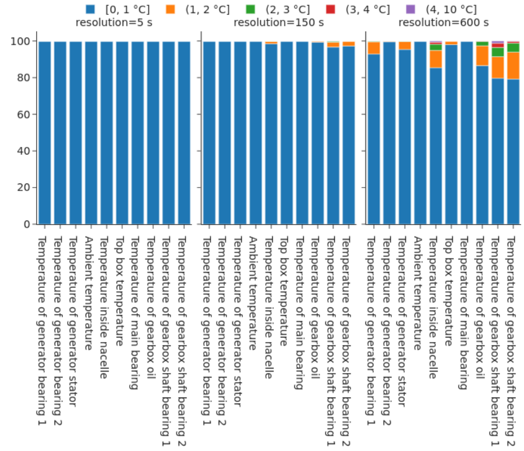


Figure 6. Information loss of three exemplary temperature signals for selected aggregation resolutions. Error bins are temperature intervals measured in degrees Celsius.

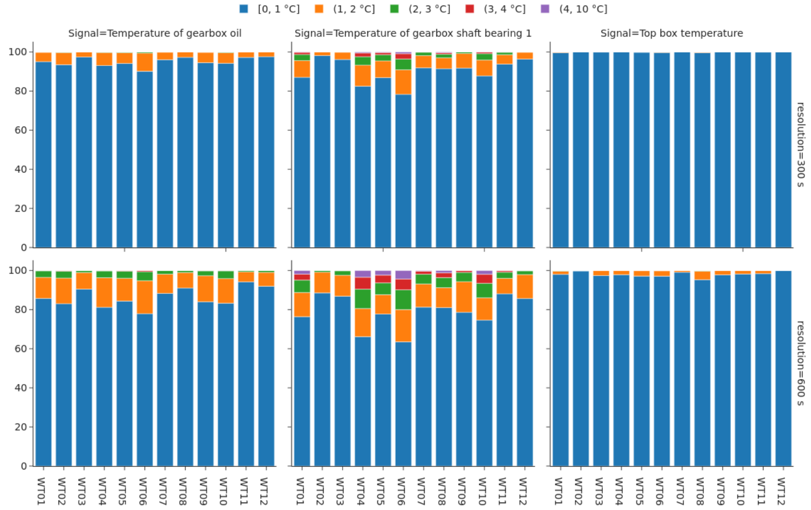


Figure 7. Information loss results subdivided into all wind turbines for the temperatures of Figure 6. Two time resolutions are provided: 300 s in the top row, 600 s in the bottom row. Error bins are temperature intervals measured in degrees Celsius.

Seasonality-Dependency

Furthermore, the results could be divided into subgroups of the month of the year, as the information loss behavior may vary along the seasons due to environmental conditions such as the wind. Figure 8 shows the variation of information loss for temperature signals throughout the year. Each month was assigned a bar, error bins and aggregation resolutions were set as in previous figures. Results for signals and resolutions not included in the figure are provided in the Supplementary Materials. The results show:

- A seasonal variation in the amount of lost information was visible for gearbox oil and gearbox shaft bearings temperatures. For the 600 s aggregation period, the percentage of data having an error below 1 °C decreased by approximately 10% between summer and winter months.
- The highest losses were registered during the months of November, December, and January when the percentage of data below 1 °C error was around 70% and 80% for the gearbox shaft bearing 1 and gearbox oil temperature respectively.
- It must be pointed out that the temperature of the gearbox oil and of the gearbox shaft bearings had a non-negligible error above 2 °C that slightly increased during the winter months.
- This seasonal dependence was also present for resolutions of 300 s, 150 s and, partly, 60 s. Due to the very low overall error, it was no longer visible for higher resolutions.
- The rest of the temperature signals were much less affected by seasonality and information loss in general. No clear differences between summer and winter months were seen in our analysis.

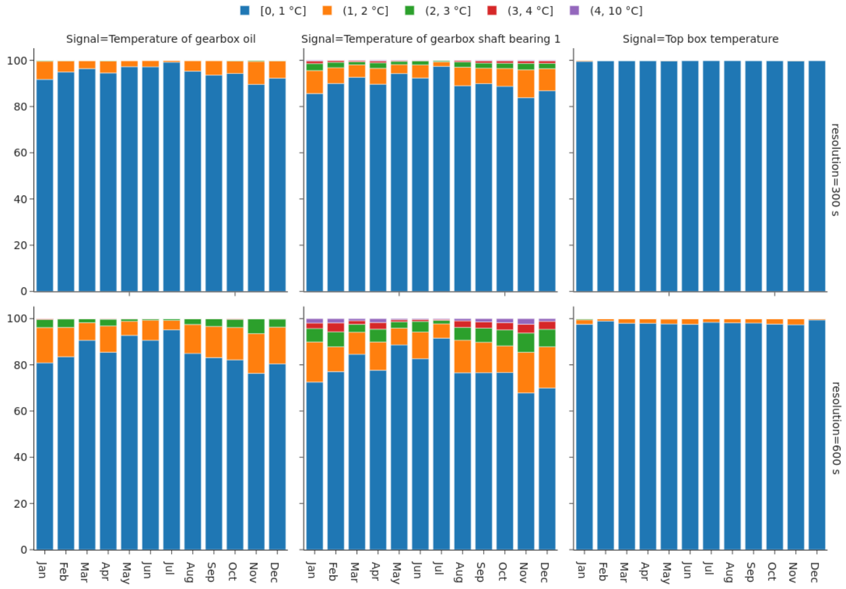


Figure 8. Information loss results for all turbines partitioned by months. Two time resolutions are provided: 300 s in the top row, 600 s in the bottom row. Error bins are temperature intervals measured in degrees Celsius.

Discussion

Within the scope of our local error approximation, wind measurements and electrical signals are largely affected by a loss of information both at short and long aggregation periods. A certain cause for this behavior is the high variability of these signals. For temperatures, on the other hand, much less error is induced by aggregation. For example, the temperature of the main bearing has more than 90% of the data with an information loss between 0 and 0.025 IQR. Thermal inertia most likely play an important role, such that the data already undergo an intrinsic reduction of the dynamics. Within all investigated temperature signals, the gearbox and generator sensors are noticeable exceptions, as they show a substantial loss of information having less than 50% of the data within the smallest error bins for 10 min aggregation. Figure 5 not only helps to quantify information loss within different signals, but also determines which measurements are most critical and thus require shorter aggregation periods.

The information loss phenomenon is relevant since it affects all turbines, as Figure 7 shows. Some turbines are affected more than others (WT04 and WT06 in the example), but overall all turbines show information loss. Moreover, knowing that some turbines are more affected than the others can be useful for modeling the behavior of the whole wind farm. In fact, these differences could indicate an existence of diverse behaviors within the turbines. However, these strong differences between the turbines within our data might also boil down to our randomly chosen sample size of 1000 h resulting in merely an average of 84 h per turbine.

It is further observed that information loss referred to transmission related temperatures is affected by seasonality. Figure 8 shows signs of such effects. For the gearbox oil and gearbox shaft bearing temperatures, an increase of 10% in the data of the [0–1 °C] error bins of the winter months with respect to the summer could be related to the variability of environmental conditions. Lower and stable wind speeds lead to lower variations in operating conditions and, consequently, less dynamics in the form of steep gradients in the temperatures of the gearbox. This results in less differences between the original and aggregated signal. Similarly, differences in external temperatures might increase or decrease the gradients of component temperatures. Figure 9 shows the wind speed and ambient temperature along the year for an exemplary turbine WT05. Note that almost identical profiles are observed for the rest of the farm. As expected, summer months are warmer, but also the range of variation as well as the average value of wind speed is lower when compared to winter. This supports the assumption of an influence of the environmental variability on an increase in information loss.

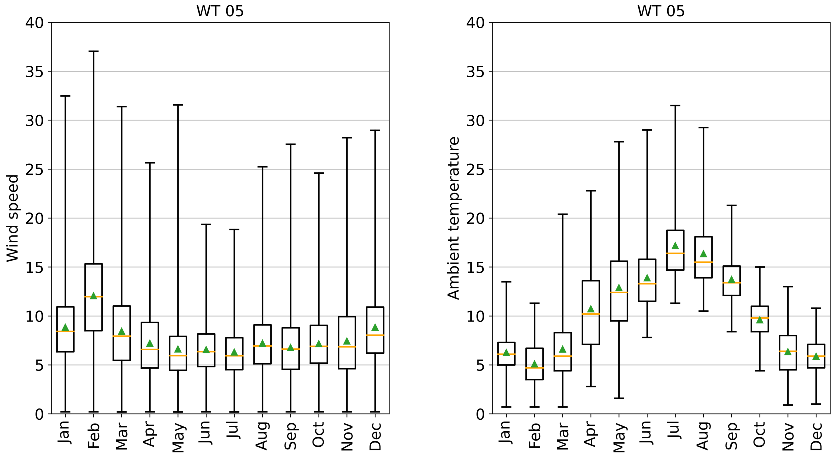


Figure 9. Boxplots representing wind speed (**left**) and ambient temperature (**right**) of an exemplary turbine (WT05) along the year.

In conclusion, the choice of the aggregation period for each signal is subjected to a benefit-costs analysis in which the ultimate use of the signal as well as data storage and transmission costs play a contrasting role. However, to detect pre-failure states higher resolution might be needed as anomalies could manifest themselves on shorter timescales.

With all these considerations, wind related measurements should be stored at the highest frequency possible since their dynamics is particularly fast and the lost information are useful for any fine-grained analyses that takes the wind behavior into consideration. A similar recommendation can be given for electrical signals that accumulate a visible amount of information losses for resolutions above 60 seconds. Temperature signals are less affected by the phenomenon of information loss upon aggregation. Consequently, they can safely be stored at lower frequencies, even though the standard SCADA resolution of 10 min is not recommendable. Concluding from the investigations, at least 150 to 300 seconds should be preferred.

4.2. Q2: Do External Conditions Have an Effect on Information Loss?

Knowing that operating conditions of turbines vary considerably both on short, i.e., hours, and long timescales, i.e., months of the year, an analysis of the relation between information loss and external conditions is presented. As wind speed is one of the most important parameters governing turbine operations, the analysis focuses specifically on this.

The objective is to characterize the behavior of the information loss with respect to the wind speed. We answer the questions whether certain wind conditions cause larger deviations in the aggregated signal and, if so, what the expected range of information loss is. Then, the behavior of different signals is compared to analyze possible shared patterns—such as certain wind speed regions—for which most signals show large variations in information loss.

Methodology

To capture the influence of wind conditions, the information loss results were divided into bins of wind speed with a size of 0.5 m/s. This value is commonly used to group data for power curve calculations and analyses of the turbine behavior [42].

Two complementary perspectives on the problem were proposed. The first attempts to capture the overall behavior of the available signals, with the objective to determine shared trends. This was accomplished by sorting the results of Equation (3) by wind speed. For each bin the mean value of information loss was computed. Like throughout the rest of the paper the IQR was used as the normalization basis. The second perspective is a detailed representation of the span of information loss per wind speed bin for each individual signal, providing an estimation of the range of variation. The range defined by the 5-95th percentile of the distribution was calculated for the different signals and for each wind speed bin. For this analysis the sign of the deviations from the original signal is relevant and must be preserved. Thus, Equation (3) was modified to Equation (4), where the direction of the difference between aggregated and raw signal is no longer omitted, resulting in a newly defined signed information loss:

$$\Lambda^{\pm}(t)_{\Delta t_{\mathrm{agg}}} = s(t)_{\Delta t_{\mathrm{agg}}} - s(t)_{\Delta t_{\mathrm{org}}} \tag{4}$$

Results

The normalized results of the mean information loss for each signal are represented in Figure 10 in the form of a heat map, such that signals can be easily compared to each other. The horizontal axis shows wind bins with a width of 0.5 m/s, the vertical axis lists the signals. The magnitude of the mean information loss for a given condition is determined by the color of the cell. Notice that the color scale is capped to a value of 0.25. Otherwise, i.e., with a full scale, the grid frequency and wind speed error values would completely mask variations in the rest of the signals. Additionally, the signals of the pitch angles are not included as they show a very large variation, augmented by the low value of their IQR. For analyzing the behavior of these signals please refer to Figure 11. Results for all temporal resolutions are available in the Supplementary Materials.

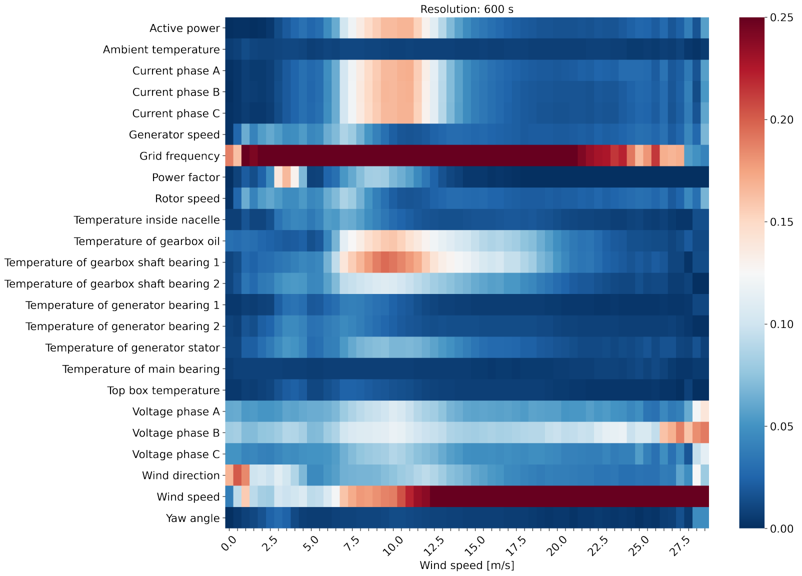


Figure 10. Heat map of the mean average of the aggregation error value per wind speed bin. Signals are normalized by the IQR and the color scale is capped to avoid grid frequency and wind speed to hide the behavior of the other signals.

From an overall perspective it can be noticed that active power, all currents, gearbox oil, and gearbox shaft bearing temperatures were characterized by a quite large mean information loss of 0.12 to 0.2 IQR for wind speeds ranging from 7 to 12 m/s. This range corresponds to the upper part load region of the power curve. Other signals follow that trend but had less pronounced maxima here such as the power factor, the temperature of the generator stator and the voltages. Albeit, the voltages showed another maximum for high wind speeds around 25 m/s. Another group of signals with the rotor and generator speed as well as the temperature within the nacelle were characterized by overall lower average losses and their maxima were located around wind speeds between 5 and 7 m/s just above the typical cut-in wind speeds of wind turbines. The frequency value had the maximum error outside the range of our heat map. However, its red band shows us that most information was lost in the operating region of the turbine from 2.5 up to 20 m/s. The wind speed itself had most of its information lost for high wind speeds. The remaining signals had a much lower span of variation of their information content. Therefore, no critical wind regimes could be easily identified. Overall, Figure 10 clearly shows that there were shared patterns in information loss between signals, but these were not unique. Some signals had larger information losses during the transition toward nominal power conditions, others were more affected at lower wind speeds, others again for high wind speeds, and finally certain signals were barely affected by changes in the range of information loss.

A second perspective focuses on the quantification of the the expected range of variation of information loss. Thus, the 5–95th percentile of the distribution of the signed local error and the mean value were sorted into the same wind bins of 0.5 m/s already used above for different levels of signal aggregation. The results for all available signals are represented in Figure 11. Values of information loss are reported in the native units of the signals without any normalization, allowing for an easy interpretation of the results. The plots further allowed us to visualize the typical profile of information loss with respect to the wind speed. Additionally, it was possible to quantify the extreme range of variation of the information content that can incur as a consequence of aggregating signals. Figure 11 is organized into subplots in which each individual signal has its own subfigure. The wind speed is assigned to the x-axis measured in m/s, on the y-axis the values of the signals in their original units are reported. Positive values indicate that the aggregated value is above the raw data, negative values the inverse, respectively. The dashed lines correspond to the values of the 5–95th percentile for each aggregation level in different colors.

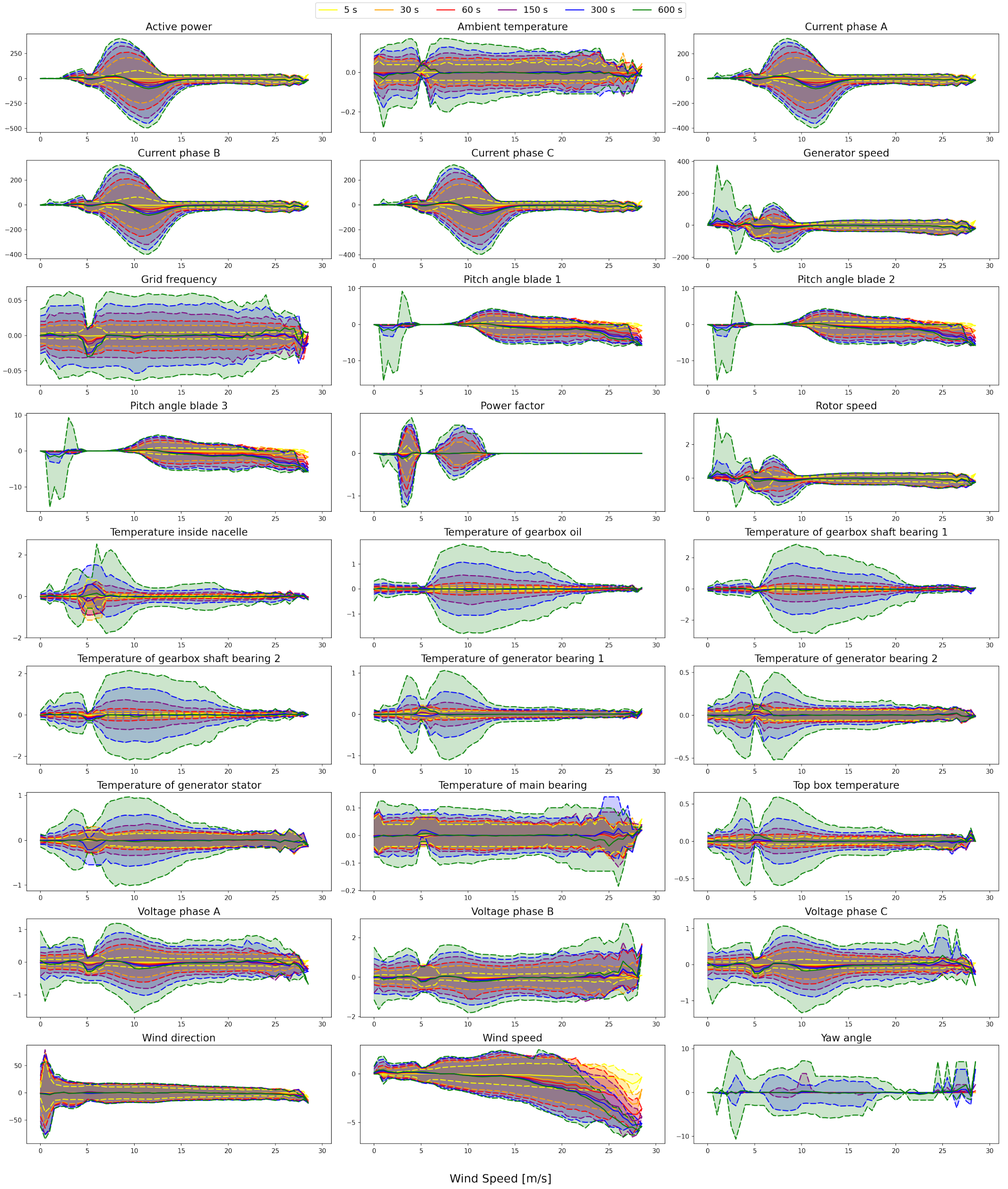


Figure 11. Line plots of the range of variation of information loss for the available signals over wind bins of 0.5 m/s width. The lower and upper dash lines represent the 5–95th percentile of the distribution. Solid lines represent mean values. The different colors denote the various aggregation periods that have been analyzed. All signals are not normalized, the y-axes are in natural units. For a less overloaded presentation, the axes do not feature any unit labelling. Please refer to Table 4 for the units of this plot.

The first observation to Figure 11 is that range of the variations in information loss always increased with the aggregation time. Some signals, in particular grid frequency, ambient, and main bearing temperature, showed little variation over the entire range of wind speed values, the span of the 5–95th range was almost constant for all wind conditions. All other signals had visible variations in their information loss ranges and their mean values oscillated around zero. As it can be seen from the dashed lines, there were some instances for which the difference range between the original and aggregated signal was particularly large, even pronounced peaks could be obtained: Those peaks can either be symmetric around 0, e.g., for all temperatures, or asymmetric, as was the case for the active power, the pitch angles, the currents, and also slightly for the voltages.

The error range was prominently large for the active power for wind speeds between 6 and 12 m/s and aggregations of 300 to 600 s with an information loss between −400 kW and 400 kW. Currents also varied heavily in that wind speed region with an information loss span between −400 and 300 A. Additionally, in the same region, temperatures had their maximal variations, in the case of the gearbox shaft bearings as high as −3 to 3 °C. However, little to no variation is seen for the main bearing temperature whose range of variation is between −0.20 and 0.35 °C. For signals such as rotor, generator speed, and the temperature of the generator bearing the range of variation of information loss was high for lower wind speeds. It was greatly reduced and constant once at nominal operating conditions with wind speed above 12 m/s. The highest error range could be observed for very low wind speeds below 5 m/s. Additionally, for pitch angles the error is extremely high with −15 to 10 deg for these low wind speeds. A similar trend of an almost constant error range for wind speeds above 12 m/s could be seen in most of the temperatures. Though, especially for temperatures related to the gearbox the transition to this constant regime was much smoother and the span of the range approached low values only for very high wind speeds, i.e., winds above 20 m/s. While most signals showed large variations around the transition phase towards nominal power, a noticeable exception was the wind direction that varies greatly between −75 and 75 degrees for wind speeds around 0 m/s and stabilized between −25 and 25 degrees for wind speeds above 5 m/s. A further prominent observation was the error of the pitch angles for wind speeds between 5 and 10 m/s where it stayed almost at 0.

The error of the wind speed itself increased steadily with increasing wind speed. Furthermore, the mean value of the error decreased, meaning an underestimation of the wind speed upon aggregation.

As a general additional remark: For aggregations up to 60 s the variation of information was low compared to aggregations of 300 to 600 s. As an example, in the case of the gearbox shaft bearing 1 temperature, information loss spanned well below −1 and 1 °C for aggregations up to 60 s, whereas for 600 s aggregation this range varied between −3 and 3 °C.

Discussion

The two analyses show that for the wind speed dependence of the information loss there is no shared pattern between all investigated signals. Nevertheless, we identified regions where maxima of the error due to aggregation can be located within the signals investigated: Below the operational regime, i.e., <5 m/s, in the transition state of the turbine from around 6 m/s to 12 m/s, and in the cut-off region above 25 m/s.

In the first region, the turbine is typically either in idle state, turned off, just starting up, or right after shutting down. The error in the generator and rotor speed most likely is caused by this start-up/shut-down transition. The asymmetry of the error in Figure 11 supports this assumption as the error for lower values of <3 m/s is positive, i.e., the aggregation value is higher than the raw value. This behavior is the consequence of the rotational speed going down. In this particular case, also the pitch angle might be part of the transition to an idle state as the asymmetric behavior is the same with the opposite sign.

The complete shutdown sequence might, however, contain important information about the health of the system that is lost by aggregation.

The second region, corresponding to the part load region of the power curve is critical for various signals. As this is a transition phase from idle condition to the full power regime, the turbine behavior is highly dynamic, leading to short term variations in the operational data. Therefore, aggregated and original signals diverge considerably. The asymmetry of roughly 5 m/s in the error span and the mean of the active power, currents, and also voltages in Figure 11 shows a general overestimation of the raw data values around 7 m/s and an underestimation for higher values around 12 m/s.

In contrast, for signals like main bearing, top box, and ambient temperature as well as the yaw angle there are no regions with large information losses. The range of variation is almost constant along the whole wind speed range. The span curves of Figure 11 might be misleading here: The error ranges are mainly below 1 °C. For a control parameter like the yaw angle this might be due to the fact that it is not directly connected to any operational state. Nevertheless, it is quite surprising that even some temperature signals of bearings do not exhibit a maximum in this dynamic transition region. One reason is most probably the thermal inertia of the material, especially for large bearings.

There is a further group of signals with the voltages that have a higher information loss also for very high wind speeds, i.e., above 25 m/s. Here, the cut-out process of the turbines could be the main influence. In the same region also the yaw angle has an error maximum. Information about the cut-out process will, therefore, be lost, when aggregating the data in low resolutions. Again other signals, such as wind direction, rotor and generator speed, and the temperature inside the nacelle have their highest errors for low wind speeds, i.e., below 5 m/s. In this region the wind direction changes more often. The errors in the generator and rotor speed most likely result from the turbine turning up or down. However, this might contain important information about the health of the system. The wind speed itself carries increased error with increasing wind speed. Its simultaneous increase in underestimation by aggregation is most likely caused by sudden bursts of wind that barely contribute to an aggregated mean value.

These observations provide useful insights on the behavior of turbine signals under specific wind conditions. In particular, they show that the accuracy of aggregated measurements is not independent from wind conditions. Gearbox behavior, for example, can vary visibly within the part load region of the power curve. Higher frequency of the data would be more appropriate to monitor and characterize these operating conditions. Figure 11 complements the analysis providing a quantification of the information loss for the different signals at various time resolutions.

The two complementary perspectives allowed to determine critical conditions during turbine operations. Various signals show a large fraction of the aggregation error concentrated for the part load and upper part load regions of the power curve. Moreover, while the aggregation error might be negligible for some signals, such as the main bearing and ambient temperature, for others, in particular active power, currents, and gearbox related temperatures, this error must not be ignored. Decreasing the length of the aggregation period to a value between 60 to 150 s greatly helps reducing the maximum extent of information loss range, maintaining a limited discrepancy between aggregated and raw signal. Moreover, knowing the link between wind speed and error provides relevant knowledge for improving the design of models aiming to describe turbine behavior.

4.3. Q3: What Is the Recommended Aggregation Frequency?

To find the optimal trade-off between minimizing the data footprint and preserving enough information to model and assess the turbine behavior, it is necessary to study the relation between information content and aggregation frequency. By knowing the behavior of an information loss over resolution it is possible to determine the critical aggregation time for a given signal, after which a great part of the information is inevitably lost. Thus, these information can be used to chose a suitable data storage solution.

Methodology

To address this relevant question the following methodology was used, that is also summarized in Algorithm 1.

Algorithm 1: Determination of the maximum aggregation time allowed for a prechosen tolerable information loss of a signal.

```
Data   : $s(t)$—Signal values time series
Input  : $\Lambda_{\text{max}}$—Tolerable information loss
         $P_{\text{min}}$—Minimum amount of $\Lambda_{\text{max}}$ in data
         $T_{\text{agg}}$—List of possible aggregation times
Result : $P[\Delta t]$—Information loss amount per aggregation time $\Delta t$
         $\Delta t_{\text{max}}$—Maximum aggregation time allowed
begin
    for $\Delta t$ in $T_{\text{agg}}$ do
        Calculate $\Lambda(t)_{\Delta t}$ from data $s(t)$;
        $n_{\text{all}}$ := count all timestamps $t$ in $\Lambda(t)_{\Delta t}$;
        $n_{\text{tolerable}}$ := count all timestamps $t$ in $\Lambda(t)_{\Delta t}$ where $\Lambda(t)_{\Delta t} \leq \Lambda_{\text{max}}$;
        $P[\Delta t] = n_{\text{tolerable}}/n_{\text{all}}$;
    end
    $\Delta t_{\text{max}} = \max \Delta t$ in $P[\Delta t]$ where $P[\Delta t] \leq P_{\text{min}}$;
end
```

First, a maximum tolerable error Λ_{max} and a minimum amount of data points P_{min} with an error lower than this error was defined. As we wanted to compare our results to each other, we used IQR normalized signal values. Therefore, Λ_{max} must be given in a percentage of the IQR. Of course, if only one signal was investigated this tolerable information loss could also be defined in real units.

Then, for each signal $s(t)$ and resolution Δt out of a list of possible aggregation resolutions T_{agg}, the information loss $\Lambda(t)_{\Delta t}$ was calculated as defined in Equation (3). This allowed us to determine the percentage $P[\Delta t]$ of data having an aggregation error lower than the chosen threshold. Thereafter, the maximum aggregation time Δt_{max} could be derived by finding the highest resolution possible for $P[\Delta t] \leq P_{\text{min}}$. Of course, the choice of the maximum error threshold had to take into consideration the final use of the data as well as the marginal cost of storing additional information.

Results

For Figure 12 we chose various tolerable error thresholds that define our closed error-bins $[0, \Lambda_{\text{max}}]$, see legend, and analyzed the available signals for a resolution range varying from 0 to 600 seconds. The horizontal axis shows the time resolution in seconds. On the vertical axis the percentage $P[\Delta t]$ of data having an aggregation error lower than the error threshold is represented. We further chose a reasonable minimal amount of $P_{\text{min}} = 80\%$ for further evaluation of the curves, indicated by a dashed line. As the figures do not share the same scale on the y-axis, it also facilitates a better orientation. Numerical results are available in the Supplementary Materials. The resulting curves show the information loss amount over resolution and lead to the following observations:

- There existed different behaviors depending on the nature of the signals. Some signals followed an "elbow curve" trend, whereas others—these include mainly temperatures—showed a more linear decay or no decay at all in information loss.
- The steepest drops were associated with wind speed and its direction, such that even short aggregation periods had less than 50% of the data below the 0.1 and 0.025 IQR error threshold respectively. Electrical signals, grid frequency in particular, also had clear drops in accordance with the results obtained in Section 4.1.3.

- Within the temperature signals, the ones related to the gearbox showed significant information losses at 600 s resolution. More than 25 to 30% of the data were above the 0.1 IQR error threshold.
- The inflexion point, i.e., the point of the strongest change in the slope, of the elbow curves indicates the aggregation period above which most of the information in the signal was lost. It can be seen that for most signals this inflexion point, for this specific error threshold, lay between 100 and 200 s.
- The higher the tolerable error was, the longer the optimal aggregation periods size could be, as the inflexion point moved towards larger values. This was explicitly shown by the rotor speed and grid frequency signals.

Discussion

Figure 12 is a proposed method for choosing the ideal resolution for turbine signals. It allows to determine the optimal resolution of a signal with the definition of a maximum error Λ_{max} and the minimum percentage P_{min} of data that should not exceed this limit. The inflexion point that is visible for most curves determines the aggregation period, above which most of the information and details of the dynamic of the signals are lost. This allows to determine a sweet spot for SCADA data storage, allowing to reduce memory footprint of the data without excessive compromises on data quality.

Moreover, the comparison of the profiles of the various curves allows to determine differences in signal dynamics. Signals with inflexion points at low aggregation periods are characterized by faster dynamics. Wind speed, wind direction, and grid frequency have a drop for aggregation periods below 10 to 100 s, after which the rate of information loss with respect to the length of the aggregation period remains almost constant and greatly reduced. Accordingly, most short-term information is contained on very short timescales. Voltage and current measurements have their inflexion point at lower resolutions between 100 and 300 s, depending on the threshold set for the maximum tolerable aggregation error. Other signals, namely temperatures have a different behavior. Instead of elbow curves they show a more linear decay or even no decay at all such as for the main bearing, whose percentage of data below the lowest error threshold, i.e., [0,0.025 IQR] is not lower than 95% even for a 600 s aggregation period.

The choice of the acceptable error Λ_{max} and the percentage P_{min} are highly dependent on the usage of the data as well as the economics of collecting, storing, and processing high volumes of data. Choosing a more restrictive error threshold moves the inflexion point towards lower values of the aggregation period size, but increases the amount of memory necessary to store information. Nevertheless, the proposed methodology allows to take informed decisions on the strategy to store and aggregate SCADA operating data. Moreover, insights concerning the dynamics of the different signals can be inferred by studying the profiles of the signal curves.

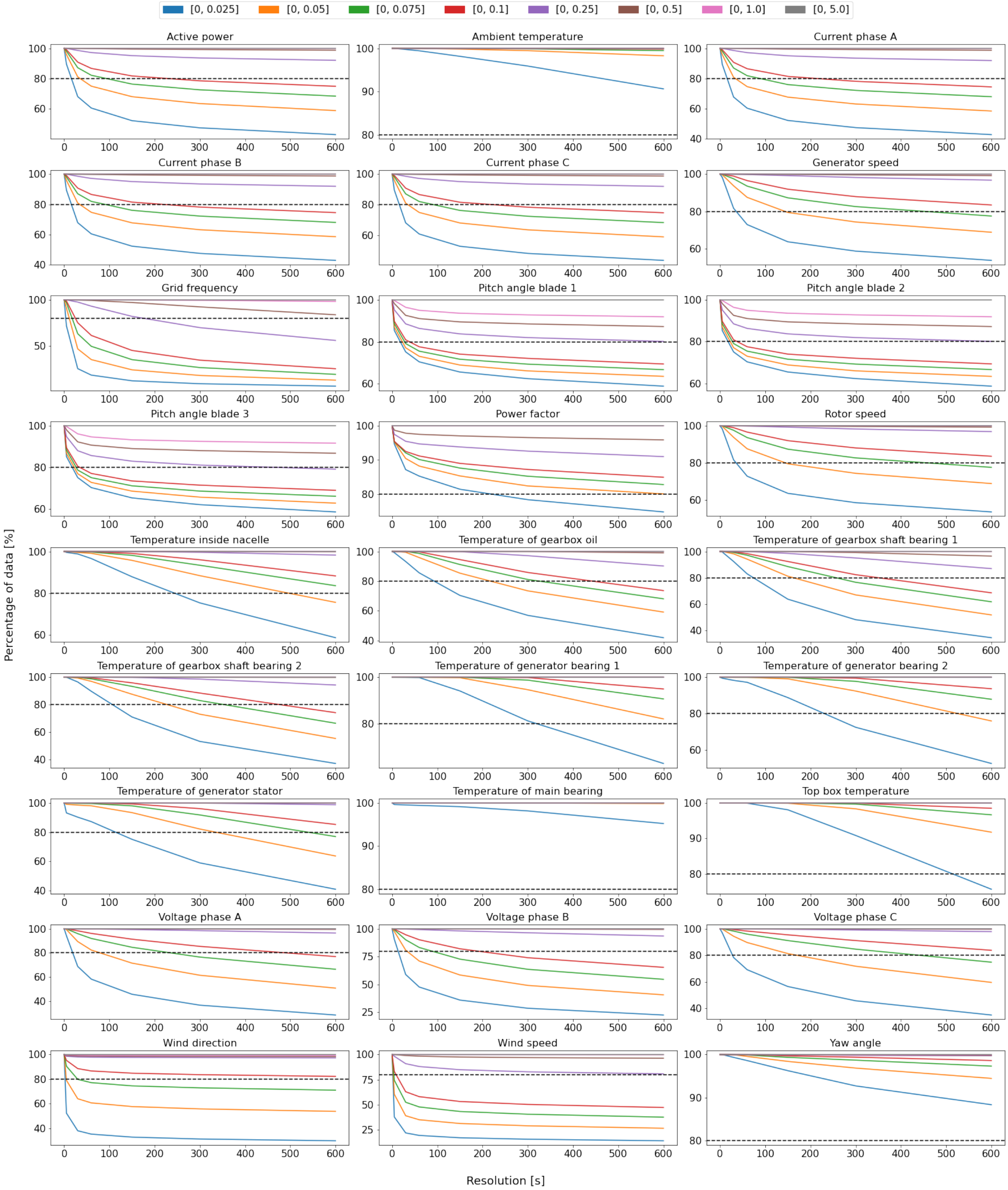


Figure 12. Relation between temporal resolution of the aggregated signal and percentage $P[\Delta t]$ of data below a given error thresholds. Multiple error limits $\Lambda_{\max}$ are represented with different colors, see legend. The scale of the y-axis is not fixed to magnify changes in signals. A black horizontal line is drawn for a value of 80% of the data to improve readability and comparison of the different plots. Labelling of the axes was omitted for a less overloaded presentation: The x-axes denote the time resolution Δt, the y-axes the data percentage P of data inside $[0, \Lambda_{\max}]$.

5. Summary and Conclusions

This study has aimed to explore the information contained in high frequency SCADA data to determine characteristics and limitations of wind turbine SCADA data. The main goal of this contribution has been to quantify the information lost due to temporal aggregation of operating data, as this data is usually only available as 10-min averaged values.

Simple methods such as the calculation of a set of descriptive statistics and the Kolmogorov–Smirnov test of the original and aggregated signal haven been carried out. Both methods, though, do not provide a clear picture of information loss. Although they show resolution-critical signals, they fail to provide any quantification of the effect of signal aggregation or indications that help to choose the optimal resolution for the signal.

To address this limitations a framework for information loss study has been elaborated. The results of this method highlight wind data and electric signals as heavily affected by information loss with less than 50% of the data with error below 2.5% of the interquartile range of the data. Temperature signals are generally less sensitive to aggregation, with the noticeable exceptions of the temperatures of the gearbox that show similar losses to the electrical signals. The presented framework allows to rank and determine the expected information for each signal and a certain aggregation period. A study of seasonal behavior has revealed that for signals measured at the gearbox, i.e., gearbox oil and gearbox shaft bearing temperatures, the information loss only varies approximately 10% between summer and winter months. Information loss is a phenomenon that affects all turbines of the analyzed wind farm, but variations in the aggregation error are seen between turbines.

Besides these approaches that pool together the whole operating regime, also the effect of wind speed on information loss has been investigated. Our study reveals that for various signals, temperatures in particular, ramping up from a stopping to rated power state causes the largest variations in the extent of information loss. Variations of an error in a 10 min aggregation interval of up to 400 kW for the active power, up to 400 A for the currents, and up to 3 °C for the main bearing temperature are the most noticeable examples of this investigation.

In addition to these considerations, a methodology to choose the optimal signal resolution is provided. To comply with stricter conditions in terms of maximum acceptable error the period of aggregation of the signal should be reduced, requiring larger resources to handle and store the signal.

In conclusion, this research delves into the limitations of typical 10-min SCADA operating data, investigates the effect of data aggregation and provides methods to determine the amount of information that is lost. Wind and electrical signals, and to a less extent temperatures of the gearbox are heavily affected by information loss and should, therefore, be stored at high resolutions of 1 to 5 s. The typical SCADA data resolution of 10 min is not sufficient to capture the dynamic behavior of these signals. The differences between the original and aggregated signal could negatively impact the performance of predictive algorithms and models describing normal turbine behaviors. Knowing the limitations of SCADA is also useful to explain the shortcomings of turbine models. Smarter SCADA data aggregation policies should be considered taking into account the issue of information losses on the various signals.

Future works on this topic can attempt to quantify information loss in terms of local minima or maxima that are lost due to aggregation. This would give further insight into information loss on the signal dynamics. An approach might be the application of Fourier transform to the data to study changes in its dynamics. However, fast changing signals, in particular wind speed, might vary too randomly, eventually making it very difficult to isolate meaningful frequencies. Still, careful filtering during analysis could lead to beneficial discoveries. For further investigations on the effects of external conditions, apart from the wind speed, it might also be important to look at other influences such as the extent of information loss under wake effects, i.e., turbulent conditions. Moreover, the case-study nature of this research does not allow to extend our conclusions to the entire universe of wind turbines, as the effect of geographical position of the wind farm, different turbine manufacturer, and technology could not be addressed.

Apart from these theoretical outlooks, it will additionally be necessary to quantify the impact of an information loss to actual analyses working with aggregated data. Although an aggregated signal might have 95% of its data with a very low information loss, the interesting operating state of a turbine could be hidden within the remaining 5% and might, therefore, be irreversibly lost. Consequently, regarding, e.g., early fault detection, a focus should be set on the question if and how early a failure is detectable with a certain resolution.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/app11178065/s1: Data Compilation S1: data for all results of Section 4.1.1 as partially displayed in Figure 2, results of the KS test of Section 4.1.2 as partly shown in Figure 3, all resulting data of Section 4.1.3 that is shown in Figure 5 and partly in Figures 6–8, the data of Figures 10 and 11 in Section 4.2 for multiple temporal resolutions, and the data of the line plots of Figure 12 in Section 4.3.

Author Contributions: conceptualization: M.B., K.P. and T.L.; methodology: M.B., K.P. and T.L.; software: M.B., K.P. and T.L.; validation: M.B., K.P. and T.L.; formal analysis: M.B., K.P. and T.L.; investigation: M.B., K.P. and T.L.; writing—original draft preparation: M.B., K.P. and T.L.; writing—review and editing: M.B. and T.L.; visualization: M.B., K.P. and T.L.; supervision: J.C. and T.L.; project administration: J.C. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the German Federal Ministry for Economic Affairs and Energy (BMWi), grant numbers 0324336A and 03EE3016B, Centro para el Desarollo Tecnológico Industrial, grant number CDTI-IDI 20191294, and Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), grant number DOCTORADO AGAUR-2017-DI 004.

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

Abbreviations

The following abbreviations are used in this manuscript:

CMS	Condition monitoring system
IQR	Interquartile range, distance between 25% and 75% of the data
KS	Kolmogorov–Smirnov
LCOE	Levelized cost of energy
O&M	Operation and maintenance
SCADA	Supervisory control and data acquisition
RUL	Residual useful lifetime

References

1. Kusiak, A.; Verma, A. Analyzing bearing faults in wind turbines: A data-mining approach. *Renew. Energy* **2012**, *48*, 110–116. [CrossRef]
2. Gonzalez, E.; Stephen, B.; Infield, D.; Melero, J.J. Using high-frequency SCADA data for wind turbine performance monitoring: A sensitivity study. *Renew. Energy* **2019**, *131*, 841–853. [CrossRef]
3. Ahmed, M.A.; Kim, Y.C. Hierarchical Communication Network Architectures for Offshore Wind Power Farms. *Energies* **2014**, *7*, 3420–3437. [CrossRef]
4. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review. *IET Renew. Power Gener.* **2016**, *11*, 382–394. [CrossRef]
5. Rohrig, K.; Berkhout, V.; Callies, D.; Durstewitz, M.; Faulstich, S.; Hahn, B.; Jung, M.; Pauscher, L.; Seibel, A.; Shan, M.; et al. Powering the 21st century by wind energy—Options, facts, figures. *Appl. Phys. Rev.* **2019**, *6*, 031303. [CrossRef]
6. Helsen, J.; Sitter, G.D.; Jordaens, P.J. Long-Term Monitoring of Wind Farms Using Big Data Approach. In Proceedings of the 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK, 29 March–1 April 2016; pp. 265–268. [CrossRef]
7. IRENA. *Renewable Capacity Statistics*; Technical Report; International Renewable Energy Agency (IRENA): Abu Dhabi, United Arab Emirates, 2021; ISBN 978-92-9260-342-7.
8. WindEurope. *Wind Energy in Europe, 2020 Statistics and the Outlook for 2021–2025*; Technical Report; WindEurope Business Intelligence: Brussels, Belgium, 2021.

9. Okumus, I.; Dinler, A. Current status of wind energy forecasting and a hybrid method for hourly predictions. *Energy Convers. Manag.* **2016**, *123*, 362–371. [CrossRef]
10. IRENA. *Renewable Power Generation Costs*; Technical Report; International Renewable Energy Agency (IRENA): Abu Dhabi, United Arab Emirates, 2020.
11. Fischer, K.; Coronado, D. Condition monitoring of wind turbines: State of the art, user experience and recommendations. *VGB PowerTech J.* **2015**, *7*, 51–56.
12. Yang, W.; Tavner, P.J.; Crabtree, C.J.; Feng, Y.; Qiu, Y. Wind turbine condition monitoring: Technical and commercial challenges: Wind turbine condition monitoring: Technical and commercial challenges. *Wind Energy* **2014**, *17*, 673–693. [CrossRef]
13. Gonzalez, E.; Stephen, B.; Infield, D.; Melero, J.J. On the use of high-frequency SCADA data for improved wind turbine performance monitoring. *J. Phys. Conf. Ser.* **2017**, *926*, 012009. [CrossRef]
14. Roberts, E.D.; Roscher, B.; Winnemöller, T.; Schelenz, R. An Investigation on the Usability of High-Frequency Wind Turbine Controller Data for Predictive Maintenance. In *Conference for Wind Power Drives 2019 : Conference Proceedings/Rik De Doncker*; RWTH Aachen University: Aachen, Germany, 2019; p. 12. [CrossRef]
15. Lin, Z.; Liu, X.; Collu, M. Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks. *Int. J. Electr. Power Energy Syst.* **2020**, *118*, 105835. [CrossRef]
16. Vargas, S.A.; Esteves, G.R.T.; Maçaira, P.M.; Bastos, B.Q.; Cyrino Oliveira, F.L.; Souza, R.C. Wind power generation: A review and a research agenda. *J. Clean. Prod.* **2019**, *218*, 850–870. [CrossRef]
17. Hanifi, S.; Liu, X.; Lin, Z.; Lotfian, S. A Critical Review of Wind Power Forecasting Methods—Past, Present and Future. *Energies* **2020**, *13*, 3764. [CrossRef]
18. Ahmadi, M.; Khashei, M. Current status of hybrid structures in wind forecasting. *Eng. Appl. Artif. Intell.* **2021**, *99*, 104133. [CrossRef]
19. Delgado, I.; Fahim, M. Wind Turbine Data Analysis and LSTM-Based Prediction in SCADA System. *Energies* **2021**, *14*, 125. [CrossRef]
20. De Felice, M.; Alessandri, A.; Ruti, P.M. Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models. *Electr. Power Syst. Res.* **2013**, *104*, 71–79. [CrossRef]
21. Jung, J.; Broadwater, R.P. Current status and future advances for wind speed and power forecasting. *Renew. Sustain. Energy Rev.* **2014**, *31*, 762–777. [CrossRef]
22. Castellani, F.; Mana, M.; Astolfi, D. An experimental analysis of wind and power fluctuations through time-resolved data of full scale wind turbines. *J. Phys. Conf. Ser.* **2018**, *1037*, 072042. [CrossRef]
23. Artigao, E.; Koukoura, S.; Honrubia-Escribano, A.; Carroll, J.; McDonald, A.; Gómez-Lázaro, E. Current Signature and Vibration Analyses to Diagnose an In-Service Wind Turbine Drive Train. *Energies* **2018**, *11*, 960. [CrossRef]
24. Siegel, D.; Zhao, W.; Lapira, E.; AbuAli, M.; Lee, J. A comparative study on vibration-based condition monitoring algorithms for wind turbine drive trains: Comparative study on wind turbine drive train health monitoring. *Wind Energy* **2014**, *17*, 695–714. [CrossRef]
25. Soua, S.; Van Lieshout, P.; Perera, A.; Gan, T.H.; Bridge, B. Determination of the combined vibrational and acoustic emission signature of a wind turbine gearbox and generator shaft in service as a pre-requisite for effective condition monitoring. *Renew. Energy* **2013**, *51*, 175–181. [CrossRef]
26. Ferrando Chacon, J.L.; Andicoberry, E.A.; Kappatos, V.; Papaelias, M.; Selcuk, C.; Gan, T.H. An experimental study on the applicability of acoustic emission for wind turbine gearbox health diagnosis. *J. Low Freq. Noise Vib. Act. Control* **2016**, *35*, 64–76. [CrossRef]
27. Inturi, V.; Sabareesh, G.; Supradeepan, K.; Penumakala, P. Integrated condition monitoring scheme for bearing fault diagnosis of a wind turbine gearbox. *J. Vib. Control* **2019**, *25*, 1852–1865. Publisher: SAGE Publications Ltd STM. [CrossRef]
28. Shokrzadeh, S.; Jafari Jozani, M.; Bibeau, E. Wind Turbine Power Curve Modeling Using Advanced Parametric and Nonparametric Methods. *IEEE Trans. Sustain. Energy* **2014**, *5*, 1262–1269. [CrossRef]
29. Pandit, R.K.; Infield, D. SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes. *IET Renew. Power Gener.* **2018**, *12*, 1249–1255. [CrossRef]
30. Schlechtingen, M.; Ferreira Santos, I. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mech. Syst.Signal Process.* **2011**, *25*, 1849–1875. [CrossRef]
31. Wilkinson, M.; Harman, K.; van Delft, T.; Darnell, B. Comparison of methods for wind turbine condition monitoring with SCADA data. *IET Renew. Power Gener.* **2014**, *8*, 390–397. [CrossRef]
32. Lutz, M.A.; Vogt, S.; Berkhout, V.; Faulstich, S.; Dienst, S.; Steinmetz, U.; Gück, C.; Ortega, A. Evaluation of Anomaly Detection of an Autoencoder Based on Maintenace Information and Scada-Data. *Energies* **2020**, *13*, 1063. [CrossRef]
33. Beretta, M.; Cárdenas, J.J.; Koch, C.; Cusidó, J. Wind Fleet Generator Fault Detection via SCADA Alarms and Autoencoders. *Appl. Sci.* **2020**, *10*, 8649. [CrossRef]
34. Beretta, M.; Julian, A.; Sepulveda, J.; Cusidó, J.; Porro, O. An Ensemble Learning Solution for Predictive Maintenance of Wind Turbines Main Bearing. *Sensors* **2021**, *21*, 1512. [CrossRef] [PubMed]
35. Alvarez, E.J.; Ribaric, A.P. An improved-accuracy method for fatigue load analysis of wind turbine gearbox based on SCADA. *Renew. Energy* **2018**, *115*, 391–399. [CrossRef]
36. Verstraeten, T.; Nowe, A.; Keller, J.; Guo, Y.; Sheng, S.; Helsen, J. Fleetwide data-enabled reliability improvement of wind turbines. *Renew. Sustain. Energy Rev.* **2019**, *109*, 428–437. [CrossRef]

37. Rott, A.; Petrović, V.; Kühn, M. Wind farm flow reconstruction and prediction from high frequency SCADA Data. *J. Phys. Conf. Ser.* **2020**, *1618*, 062067. [CrossRef]
38. Liu, J.; Khattak, A.; Han, L.; Yuan, Q. How much information is lost when sampling driving behavior data? Indicators to quantify the extent of information loss. *J. Intell. Connect. Veh.* **2020**, *3*, 17–29. [CrossRef]
39. Montgomery, D.C.; Runger, G.C. *Applied Statistics and Probability for Engineers*, 3rd ed.; Wiley: New York, NY, USA, 2003.
40. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley Series in Behavioral Science; Addison-Wesley Pub. Co.: Reading, MA, USA, 1977.
41. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed.; Wiley Series in Probability and Statistics. Applied Probability and Statistics Section; Wiley: New York, NY, USA, 1999.
42. Yang, W.; Court, R.; Jiang, J. Wind turbine condition monitoring by the approach of SCADA data analysis. *Renew. Energy* **2013**, *53*, 365–376. [CrossRef]

6. Conclusions

6.1 Summary

The wind energy industry is living a rapid and growing expansion thanks to the recent improvements in the technology which allowed to drive down LCOE, and the strong resolution of Governments and energy companies to move towards a renewable and sustainable energy sector. Nonetheless, important challenges regarding turbine maintenance and monitoring are still open. Considering the upward trend of offshore wind installation, optimized maintenance strategies become even more important, as logistics and organization of maintenance intervention are especially complicated in the open–sea.

The last decade has seen a growing demand of data–based predictive strategies in the wind energy field. Both Academia and the industry have realized that scheduled periodic checks of turbines are not enough. The harsh operating conditions, the innate variability of the wind resource, more demanding grid codes (and associated fines), and the costs caused by unexpected failures are all incentives to invest in continuous monitoring and optimization of maintenance logistics.

As presented in **Chapter** 2 multiple sources of information are available to monitor wind turbines. A noticeable trend in recent years is the rise in the utilization of SCADA data as the basis for monitoring systems. This thesis aims to design predictive maintenance strategies that are good fits for the necessity of the wind energy industry. The important characteristics of effective maintenance are captured by the following criteria.

Intepretability is important to ensure that predictions turns into actionable insights thus driving better maintenance. **Scalability** is needed due to the wide diversity of technologies available in the wind energy field. Multiple manufacturers and models are available on the market and it is common for a windfarm operator to have more than one brand of turbines at their disposal. **Modularity** is required due to the complexity of the failure patterns in wind turbines. Being very complex electro–mechanical assemble, wind turbines require a flexible and extensible solutions to capture different failure signatures. **Reliability** of predictions is needed to minimize maintanence costs, the number of false negatives and false positives must be kept low in order to better schedule maintenance and minimize expenses. Finally, the **availability of data** is a crucial point in the implementation of predictive maintenance in the industry. Vibration, acoustics and current signature sensors are hardly justifiable —due to their high costs— for turbines approaching the end of their lives. SCADA data instead is a much more compelling solutions as it does not require installation of additional sensors.

The literature review presented in **Chapter** 2 analyzed the different type of data and models that have been used in other researches. The following shortcomings have been identified:

1. *Incomplete utilization of the data available and rare mix of multiple sources of information.*
2. *Results are typically obtained on small datasets or using simulated data.*
3. *Limited availability of results tested on multiple turbine technologies and diverse operating conditions.*
4. *Lack of materials addressing qualitatively and quantitatively the limitations of SCADA data.*

All these problems are addressed in this thesis. The developed predictive frameworks attempts to make the best use of the available data and provide robust solutions for the industry by validating results on large and diverse datasets.

In **Chapter** 3 the advantages of using ensemble methods to combine different sources of information —SCADA and alarm data— is demonstrated. Information fusion is not a new topic, in fact it is well known and widely applied in other contexts, but it is rare to find wind turbine application combining multiple sources of information. Combining alarms and SCADA data is a relatively easy solution to improve predicting performances of algorithms. It can lead to more interpretable results, especially when compared to solutions based solely on complicated algorithms focusing on a single data source.

The potential of ensemble methods is further explored in **Chapter** 4. Two publications using ensembles to combine various algorithms —fitted on the same set of data— are presented. The appeal of ensembles is justified by their flexible and extensible nature, it is easy to include additional algorithms capturing different characteristics of the data and balancing interpretability against predicting performances.

Moreover, multiple predictive paradigms have been investigated in **Chapter** 4. Namely, information has been extracted by monitoring the behavior of turbines through times, but also making comparisons within turbines in the same windfarm. The constantly retrained normality model used in 4.2 is an example of a solution which is able to detect trends in time series data, and gradual degradation of components' conditions. Whereas, anomaly detection models and analysis of statistical indicators are good options for detecting differences within turbines in the same windfarm. Thus, spotting those turbines that behaves differently from the rest. These are complementary approaches to tackle the same problem; Ensemble methods allow to benefit from the advantages and limit the shortcomings of each individual approach.

All the mentioned results have been proven on real datasets coming from multiple windfarms composed of a vast selections of the most common turbine brands on the market (Vestas, Acciona, Siemens Gamesa, Nordex, Senvion, Enercon). The performance results have been obtained tracking and evaluating metrics on multiple years of operation, in order to provide a faithful and reliable estimation of prediction performances. Moreover, these solutions have been included in SMARTIVE offering and tested on pilot and commercial projects with large energy industry companies. Algorithms have been utilized to monitor entire windfarm fleet composed by hundreds of assets located all around the world.

Another important aspect is the attention towards scalable and automatic predicting solutions. In particular, unsupervised models and algorithms characterized by limited preprocessing, are preferred to supervised approaches —such as fault classifiers. This decision was driven by the burden that data labeling and preprocessing imposes on model scalability. Supervised models can be a good solution in presence of reliable and standardized events data —which can be used to assign labels and identify failures. Unfortunately, having good work order logs is rare in the industry. Events logs, unlike SCADA data rarely follow standardized guidelines. Logs may vary a lot between different manufacturer, operators, and sometimes even windfarms. The formats of the data is rarely designed using machine–readable standards. Finally, the predictions of a classifier are not always easy to interpret.

The last important contribution of this work regards the discussion on the limitations of SCADA data that is presented in **Chapter** 5. Other authors have discussed and documented the shortcomings of SCADA, but very few quantitative estimations are available. The presented article developed a quantitative framework to assess the amount of information that is lost due to SCADA data aggregation. Moreover, this framework was used to determine the impact of wind speed on the the information loss, and characterize the relation between aggregating resolution and information content. On one hand, the article shows that the typical time resolution of SCADA data (i.e.

5-10 minutes) is not able to characterize fast changing signals, such as wind–related or electrical measurements. In these cases logging the minimum or maximum value within the aggregation period can be useful to characterize relevant aspects of the signal. The framework that has been developed can be a useful tool for windfarm operator to take better decisions when it comes to data storage policies. Considering the growing importance that SCADA data is taking in the predictive maintenance field, the balance between data footprint and richness of the information should be reconsidered.

6.2 Suggestions for Future Works

Various lines of investigations can be built upon the results of this thesis, the most significant are mentioned in the following paragraphs.

- Information from work order logs, oil analyses, etc. can be incorporated to the predictive framework presented in **Chapter** 3. This data sources might provide additional complementary information to assess the status of turbine components.
- Alternative ranking schemes and ensemble methods can be investigated. In this work simple approaches have been proposed but proportional weights to indicators can be assigned based on prior expert knowledge. Alternatively, in the presence of reliable labels it could be useful to train a meta–classifier assigning weights to base indicators based on historical performances.
- The proposed frameworks can be used to assess the status of other turbine components. In particular, the study of the transformer and pitch system is recommended giving the importance of both system in turbines' operations.
- The information regarding the health status of a turbine can be fed to fault-tolerant controlling algorithms. In this way, the asset may be maintained in operation at lower load conditions, without further compromising its conditions. This would be particularly interesting considering the amount of time that is often needed to schedule and complete corrective actions on main components of turbines.
- The investigation regarding limitations of SCADA data can be expanded, taking into account details that have been neglected in 5.2. The effect of environmental conditions and different turbine technologies can be studied determining the sensitivity of information loss to these parameters. Another compelling question might be the impact of information

loss on the capability to assess turbine conditions. In other words, coarse aggregation of SCADA data might be limiting the capability of detecting failures early in wind turbines?

Bibliography

[1] International Energy Agency, *Renewable Energy Market Update: Outlook for 2020 and 2021*, en. OECD, Jun. 2020, ISBN: 978-92-64-48868-7. DOI: 10.1787/afbc8c1d-en. [Online]. Available: https://www.oecd-ilibrary.org/energy/renewable-energy-market-update_afbc8c1d-en (visited on 11/14/2021).

[2] M. Wolsink, “Wind power implementation: The nature of public attitudes: Equity and fairness instead of ‘backyard motives’,” en, *Renewable and Sustainable Energy Reviews*, vol. 11, no. 6, pp. 1188–1207, Aug. 2007, ISSN: 13640321. DOI: 10.1016/j.rser.2005.10.005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032105001255 (visited on 11/14/2021).

[3] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, “A systematic literature review of machine learning methods applied to predictive maintenance,” en, *Computers & Industrial Engineering*, vol. 137, p. 106 024, Nov. 2019, ISSN: 0360-8352. DOI: 10.1016/j.cie.2019.106024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360835219304838 (visited on 11/14/2021).

[4] B. Hahn, M. Durstewitz, and K. Rohrig, “Reliability of Wind Turbines,” en, in *Wind Energy*, J. Peinke, P. Schaumann, and S. Barth, Eds., Berlin, Heidelberg: Springer, 2007, pp. 329–332, ISBN: 978-3-540-33866-6. DOI: 10.1007/978-3-540-33866-6_62.

[5] L. Ziegler, E. Gonzalez, T. Rubert, U. Smolka, and J. J. Melero, “Lifetime extension of onshore wind turbines: A review covering Germany, Spain, Denmark, and the UK,” en, *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1261–1271, Feb. 2018, ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.09.100. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032117313503 (visited on 11/21/2021).

[6] J. L. Ferrando Chacon, E. A. Andicoberry, V. Kappatos, M. Papaelias, C. Selcuk, and T.-H. Gan, “An experimental study on the applicability of acoustic emission for wind turbine gear-

box health diagnosis," en, *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 35, no. 1, pp. 64–76, Mar. 2016, Publisher: SAGE Publications Ltd STM, ISSN: 1461-3484. DOI: 10.1177/0263092316628401. [Online]. Available: https://doi.org/10.1177/0263092316628401 (visited on 11/14/2021).

[7] V. Inturi, G. Sabareesh, K. Supradeepan, and P. Penumakala, "Integrated condition monitoring scheme for bearing fault diagnosis of a wind turbine gearbox," en, *Journal of Vibration and Control*, vol. 25, no. 12, pp. 1852–1865, Jun. 2019, Publisher: SAGE Publications Ltd STM, ISSN: 1077-5463. DOI: 10.1177/1077546319841495. [Online]. Available: https://doi.org/10.1177/1077546319841495 (visited on 11/14/2021).

[8] S. Soua, P. Van Lieshout, A. Perera, T.-H. Gan, and B. Bridge, "Determination of the combined vibrational and acoustic emission signature of a wind turbine gearbox and generator shaft in service as a pre-requisite for effective condition monitoring," en, *Renewable Energy*, vol. 51, pp. 175–181, Mar. 2013, ISSN: 0960-1481. DOI: 10.1016/j.renene.2012.07.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148112004302 (visited on 11/14/2021).

[9] D. Siegel, W. Zhao, E. Lapira, M. AbuAli, and J. Lee, "A comparative study on vibration-based condition monitoring algorithms for wind turbine drive trains," *Wind Energy*, vol. 17, no. 5, pp. 695–714, 2014. DOI: https://doi.org/10.1002/we.1585. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.1585. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/we.1585.

[10] E. Artigao, S. Koukoura, A. Honrubia-Escribano, J. Carroll, A. McDonald, and E. Gómez-Lázaro, "Current Signature and Vibration Analyses to Diagnose an In-Service Wind Turbine Drive Train," en, *Energies*, vol. 11, no. 4, p. 960, Apr. 2018, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/en11040960. [Online]. Available: https://www.mdpi.com/1996-1073/11/4/960 (visited on 11/14/2021).

[11] M. A. Ahmed and Y.-C. Kim, "Hierarchical Communication Network Architectures for Offshore Wind Power Farms," en, *Energies*, vol. 7, no. 5, pp. 3420–3437, May 2014, Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/en7053420. [Online]. Available: https://www.mdpi.com/1996-1073/7/5/3420 (visited on 11/14/2021).

[12] J. Tautz-Weinert and S. J. Watson, "Using SCADA data for wind turbine condition monitoring – a review," en, *IET Renewable Power Generation*, vol. 11, no. 4, pp. 382–394, Mar. 2017, ISSN: 1752-1416, 1752-1424. DOI: 10.1049/iet-rpg.2016.0248. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1049/iet-rpg.2016.0248 (visited on 11/27/2021).

[13] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, J. Keane, and G. Nenadic, "Machine learning methods for wind turbine condition monitoring: A review," en, *Renewable Energy*, vol. 133, pp. 620–635, Apr. 2019, ISSN: 09601481. DOI: 10.1016/j.renene.2018.10.047. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S096014811831231X (visited on 11/27/2021).

[14] W. Qiao and D. Lu, "A Survey on Wind Turbine Condition Monitoring and Fault Diagnosis—Part II: Signals and Signal Processing Methods," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6546–6557, Oct. 2015, Conference Name: IEEE Transactions on Industrial Electronics, ISSN: 1557-9948. DOI: 10.1109/TIE.2015.2422394.

[15] D. Lu and W. Qiao, "Frequency demodulation-aided condition monitoring for drivetrain gearboxes," in *2013 IEEE Transportation Electrification Conference and Expo (ITEC)*, Jun. 2013, pp. 1–6. DOI: 10.1109/ITEC.2013.6574526.

[16] S. Shanbr, F. Elasha, M. Elforjani, and J. Teixeira, "Detection of natural crack in wind turbine gearbox," en, *Renewable Energy*, vol. 118, pp. 172–179, Apr. 2018, ISSN: 0960-1481. DOI: 10.1016/j.renene.2017.10.104. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148117310741 (visited on 11/22/2021).

[17] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," en, *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, Feb. 2011, ISSN: 08883270. DOI: 10.1016/j.ymssp.2010.07.017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0888327010002530 (visited on 11/23/2021).

[18] R. Randall and W. Smith, "New cepstral methods for the diagnosis of gear and bearing faults under variable speed conditions," *ICSV 2016 - 23rd International Congress on Sound and Vibration: From Ancient to Modern Acoustics*, 2016.

[19] R. B. Randall, "A history of cepstrum analysis and its application to mechanical problems," en, *Mechanical Systems and Signal Processing*, Special Issue on Surveillance, vol. 97, pp. 3–19, Dec. 2017, ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2016.12.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327016305556 (visited on 11/23/2021).

[20] A. S. Sait and Y. I. Sharaf-Eldeen, "A Review of Gearbox Condition Monitoring Based on vibration Analysis Techniques Diagnostics and Prognostics," in *Rotating Machinery, Structural Health Monitoring, Shock and Vibration, Volume 5*, T. Proulx, Ed., New York, NY: Springer New York, 2011, pp. 307–324, ISBN: 978-1-4419-9428-8.

[21] P. Santos, L. F. Villa, A. Reñones, A. Bustillo, and J. Maudes, “Wind Turbines Fault Diagnosis Using Ensemble Classifiers,” en, in *Advances in Data Mining. Applications and Theoretical Aspects*, vol. 7377, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 67–76, ISBN: 978-3-642-31487-2. DOI: 10.1007/978-3-642-31488-9_6. [Online]. Available: http://link.springer.com/10.1007/978-3-642-31488-9_6 (visited on 11/23/2021).

[22] Z. Zhang, A. Verma, and A. Kusiak, “Fault Analysis and Condition Monitoring of the Wind Turbine Gearbox,” *IEEE Transactions on Energy Conversion*, vol. 27, no. 2, pp. 526–535, Jun. 2012, Conference Name: IEEE Transactions on Energy Conversion, ISSN: 1558-0059. DOI: 10.1109/TEC.2012.2189887.

[23] G. Jiang, H. He, J. Yan, and P. Xie, “Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox,” *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019, Conference Name: IEEE Transactions on Industrial Electronics, ISSN: 1557-9948. DOI: 10.1109/TIE.2018.2844805.

[24] P. G. Lind, L. Vera-Tudela, M. Wächter, M. Kühn, and J. Peinke, “Normal Behaviour Models for Wind Turbine Vibrations: Comparison of Neural Networks and a Stochastic Approach,” en, *Energies*, vol. 10, no. 12, p. 1944, Dec. 2017, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/en10121944. [Online]. Available: https://www.mdpi.com/1996-1073/10/12/1944 (visited on 11/23/2021).

[25] S. Guo, T. Yang, H. Hua, and J. Cao, “Coupling fault diagnosis of wind turbine gearbox based on multitask parallel convolutional neural networks with overall information,” en, *Renewable Energy*, vol. 178, pp. 639–650, Nov. 2021, ISSN: 0960-1481. DOI: 10.1016/j.renene.2021.06.088. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096014812100954X (visited on 11/23/2021).

[26] A. Movsessian, D. García Cava, and D. Tcherniak, “An artificial neural network methodology for damage detection: Demonstration on an operating wind turbine blade,” en, *Mechanical Systems and Signal Processing*, vol. 159, p. 107 766, Oct. 2021, ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2021.107766. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327021001618 (visited on 11/22/2021).

[27] X. Gong and W. Qiao, “Current-Based Mechanical Fault Detection for Direct-Drive Wind Turbines via Synchronous Sampling and Impulse Detection,” *IEEE Transactions on Industrial Electronics*, vol. 62, no. 3, pp. 1693–1702, Mar. 2015, Conference Name: IEEE Transactions on Industrial Electronics, ISSN: 1557-9948. DOI: 10.1109/TIE.2014.2363440.

[28] Y. Merizalde, L. Hernández-Callejo, O. Duque-Perez, and R. A. López-Meraz, "Fault Detection of Wind Turbine Induction Generators through Current Signals and Various Signal Processing Techniques," en, *Applied Sciences*, vol. 10, no. 21, p. 7389, Jan. 2020, Number: 21 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/app10217389. [Online]. Available: https://www.mdpi.com/2076-3417/10/21/7389 (visited on 11/23/2021).

[29] W. Qiao and L. Qu, "Prognostic condition monitoring for wind turbine drivetrains via generator current analysis," *Chinese Journal of Electrical Engineering*, vol. 4, no. 3, pp. 80–89, Sep. 2018, Conference Name: Chinese Journal of Electrical Engineering, ISSN: 2096-1529. DOI: 10.23919/CJEE.2018.8471293.

[30] M. P. Barrett and J. Stover, "Understanding Oil Analysis: How It Can Improve Reliability of Wind Turbine Gearboxes," en, p. 8, [Online]. Available: https://www.geartechnology.com/articles/1113/Understanding_Oil_Analysis:_How_it_Can_Improve_Reliability_of_Wind_Turbine_Gearboxes (visited on 01/22/2022).

[31] D. Kr Singh, J. Kurien, and A. Villayamore, "Study and analysis of wind turbine gearbox lubrication failure and its mitigation process," en, *Materials Today: Proceedings*, vol. 44, pp. 3976–3983, 2021, ISSN: 22147853. DOI: 10.1016/j.matpr.2020.10.047. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214785320376136 (visited on 11/23/2021).

[32] M. A. Rodríguez-López, L. M. López-González, L. M. López-Ochoa, and J. Las-Heras-Casas, "Development of indicators for the detection of equipment malfunctions and degradation estimation based on digital signals (alarms and events) from operation SCADA," en, *Renewable Energy*, vol. 99, pp. 224–236, Dec. 2016, ISSN: 09601481. DOI: 10.1016/j.renene.2016.06.056. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148116305808 (visited on 11/23/2021).

[33] K. Leahy, C. Gallagher, P. O'Donovan, and D. T. O'Sullivan, "Cluster analysis of wind turbine alarms for characterising and classifying stoppages," *IET Renewable Power Generation*, vol. 12, no. 10, pp. 1146–1154, 2018. DOI: https://doi.org/10.1049/iet-rpg.2017.0422. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rpg.2017.0422. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2017.0422.

[34] E. Gonzalez, M. Reder, and J. J. Melero, "SCADA alarms processing for wind turbine component failure detection," en, *Journal of Physics: Conference Series*, vol. 753, p. 072 019, Sep. 2016, ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/753/7/072019. [Online].

Available: https://iopscience.iop.org/article/10.1088/1742-6596/753/7/072019 (visited on 11/23/2021).

[35] "Wind turbine SCADA alarm analysis for improving reliability," DOI: 10.1002/we.513. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/we.513 (visited on 11/27/2021).

[36] D. Hadžiosmanović, D. Bolzoni, and P. H. Hartel, "A log mining approach for process monitoring in SCADA," en, *International Journal of Information Security*, vol. 11, no. 4, pp. 231–251, Aug. 2012, ISSN: 1615-5262, 1615-5270. DOI: 10.1007/s10207-012-0163-8. [Online]. Available: http://link.springer.com/10.1007/s10207-012-0163-8 (visited on 11/23/2021).

[37] P. Urban and L. Landryová, "Identification and evaluation of alarm logs from the alarm management system," in *2016 17th International Carpathian Control Conference (ICCC)*, May 2016, pp. 769–774. DOI: 10.1109/CarpathianCC.2016.7501199.

[38] M. Wilkinson, B. Hendriks, F. Spinato, K. Harman, E. Gomez, H. Bulacio, J. Roca, P. Tavner, Y. Feng, and H. Long, "Methodology and results of the reliawind reliability field study," in *European Wind Energy Conference , EWEC 2010*, vol. 3, 2010, pp. 1984–2004. [Online]. Available: https://eprints.whiterose.ac.uk/83343/.

[39] E. Gonzalez, B. Stephen, D. Infield, and J. J. Melero, "Using high-frequency SCADA data for wind turbine performance monitoring: A sensitivity study," en, *Renewable Energy*, vol. 131, pp. 841–853, Feb. 2019, ISSN: 09601481. DOI: 10.1016/j.renene.2018.07.068. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148118308656 (visited on 11/24/2021).

[40] A. Kusiak and W. Li, "The prediction and diagnosis of wind turbine faults," en, *Renewable Energy*, vol. 36, no. 1, pp. 16–23, Jan. 2011, ISSN: 0960-1481. DOI: 10.1016/j.renene.2010.05.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148110002338 (visited on 11/24/2021).

[41] C. A. Walford, "Wind Turbine Reliability: Understanding and Minimizing Wind Turbine Operation and Maintenance Costs," Mar. 2006. DOI: 10.2172/882048. [Online]. Available: https://www.osti.gov/biblio/882048.

[42] F. Spinato, P. Tavner, G. van Bussel, and E. Koutoulakos, "Reliability of wind turbine subassemblies," en, *IET Renewable Power Generation*, vol. 3, no. 4, p. 387, 2009, ISSN: 17521416. DOI: 10.1049/iet-rpg.2008.0060. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-rpg.2008.0060 (visited on 11/24/2021).

[43] M. Schlechtingen and I. Ferreira Santos, "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection," en, *Mechanical Systems and Signal Processing*, vol. 25, no. 5, pp. 1849–1875, Jul. 2011, ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2010.12.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327010004310 (visited on 07/13/2021).

[44] W. Yang, R. Court, and J. Jiang, "Wind turbine condition monitoring by the approach of SCADA data analysis," en, *Renewable Energy*, vol. 53, pp. 365–376, May 2013, ISSN: 09601481. DOI: 10.1016/j.renene.2012.11.030. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148112007653 (visited on 11/27/2021).

[45] Y. Feng, Y. Qiu, C. Crabtree, H. Long, and P. Tavner, "Use of scada and cms signals for failure detection and diagnosis of a wind turbine gearbox," in *European Wind Energy Association Conference, EWEA 2011*, 2011, pp. 17–19. [Online]. Available: https://eprints.whiterose.ac.uk/83334/.

[46] A. Zaher, S. McArthur, D. Infield, and Y. Patel, "Online wind turbine fault detection through automated SCADA data analysis," en, *Wind Energy*, vol. 12, no. 6, pp. 574–593, 2009, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.319, ISSN: 1099-1824. DOI: 10.1002/we.319. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/we.319 (visited on 11/27/2021).

[47] J. Maldonado-Correa, S. Martín-Martínez, E. Artigao, and E. Gómez-Lázaro, "Using SCADA Data for Wind Turbine Condition Monitoring: A Systematic Literature Review," en, *Energies*, vol. 13, no. 12, p. 3132, Jun. 2020, ISSN: 1996-1073. DOI: 10.3390/en13123132. [Online]. Available: https://www.mdpi.com/1996-1073/13/12/3132 (visited on 11/27/2021).

[48] E. Gonzalez, B. Stephen, D. Infield, and J. J. Melero, "On the use of high-frequency SCADA data for improved wind turbine performance monitoring," en, *Journal of Physics: Conference Series*, vol. 926, p. 012 009, Nov. 2017, ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/926/1/012009. [Online]. Available: http://stacks.iop.org/1742-6596/926/i=1/a=012009?key=crossref.0b7db9f536b281bb1e6c90d4a1ed6fe3 (visited on 01/30/2020).

[49] A. Santolamazza, D. Dadi, and V. Introna, "A Data-Mining Approach for Wind Turbine Fault Detection Based on SCADA Data Analysis Using Artificial Neural Networks," en, *Energies*, vol. 14, no. 7, p. 1845, Mar. 2021, ISSN: 1996-1073. DOI: 10.3390/en14071845. [Online]. Available: https://www.mdpi.com/1996-1073/14/7/1845 (visited on 11/27/2021).

[50] C. Velandia-Cardenas, Y. Vidal, and F. Pozo, “Wind Turbine Fault Detection Using Highly Imbalanced Real SCADA Data,” en, *Energies*, vol. 14, no. 6, p. 1728, Mar. 2021, ISSN: 1996-1073. DOI: 10.3390/en14061728. [Online]. Available: https://www.mdpi.com/1996-1073/14/6/1728 (visited on 11/27/2021).

[51] Y. Zhao, D. Li, A. Dong, D. Kang, Q. Lv, and L. Shang, “Fault Prediction and Diagnosis of Wind Turbine Generators Using SCADA Data,” en, *Energies*, vol. 10, no. 8, p. 1210, Aug. 2017, ISSN: 1996-1073. DOI: 10.3390/en10081210. [Online]. Available: http://www.mdpi.com/1996-1073/10/8/1210 (visited on 12/06/2021).

[52] H. Rashid, E. Khalaji, J. Rasheed, and C. Batunlu, “Fault Prediction of Wind Turbine Gearbox Based on SCADA Data and Machine Learning,” in *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, Sep. 2020, pp. 391–395. DOI: 10.1109/ACIT49673.2020.9208884.

[53] F. Ding, Z. Tian, F. Zhao, and H. Xu, “An integrated approach for wind turbine gearbox fatigue life prediction considering instantaneously varying load conditions,” en, *Renewable Energy*, vol. 129, pp. 260–270, Dec. 2018, ISSN: 09601481. DOI: 10.1016/j.renene.2018.05.074. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148118305950 (visited on 11/27/2021).

[54] B. Corley, J. Carroll, and A. Mcdonald, “Fault detection of wind turbine gearbox using thermal network modelling and SCADA data,” en, *Journal of Physics: Conference Series*, vol. 1618, no. 2, p. 022 042, Sep. 2020, ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/1618/2/022042. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1618/2/022042 (visited on 11/27/2021).

[55] B. Corley, S. Koukoura, J. Carroll, and A. McDonald, “Combination of Thermal Modelling and Machine Learning Approaches for Fault Detection in Wind Turbine Gearboxes,” en, *Energies*, vol. 14, no. 5, p. 1375, Mar. 2021, ISSN: 1996-1073. DOI: 10.3390/en14051375. [Online]. Available: https://www.mdpi.com/1996-1073/14/5/1375 (visited on 11/27/2021).

[56] C. Carrillo, A. Obando Montaño, J. Cidrás, and E. Díaz-Dorado, “Review of power curve modelling for wind turbines,” en, *Renewable and Sustainable Energy Reviews*, vol. 21, pp. 572–581, May 2013, ISSN: 13640321. DOI: 10.1016/j.rser.2013.01.012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032113000439 (visited on 12/05/2021).

[57] M. Lydia, S. S. Kumar, A. I. Selvakumar, and G. E. Prem Kumar, “A comprehensive review on wind turbine power curve modeling techniques,” en, *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 452–460, Feb. 2014, ISSN: 13640321. DOI: 10.1016/j.rser.2013.

10.030. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032113007296 (visited on 12/05/2021).

[58] V. Sohoni, S. C. Gupta, and R. K. Nema, "A Critical Review on Wind Turbine Power Curve Modelling Techniques and Their Applications in Wind Based Energy Systems," en, *Journal of Energy*, vol. 2016, e8519785, Jul. 2016, Publisher: Hindawi, ISSN: 2356-735X. DOI: 10.1155/2016/8519785. [Online]. Available: https://www.hindawi.com/journals/jen/2016/8519785/ (visited on 12/05/2021).

[59] D. Astolfi, "Wind Turbine Operation Curves Modelling Techniques," en, *Electronics*, vol. 10, no. 3, p. 269, Jan. 2021, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/electronics10030269. [Online]. Available: https://www.mdpi.com/2079-9292/10/3/269 (visited on 12/05/2021).

[60] F. Pelletier, C. Masson, and A. Tahan, "Wind turbine power curve modelling using artificial neural network," en, *Renewable Energy*, vol. 89, pp. 207–214, Apr. 2016, ISSN: 09601481. DOI: 10.1016/j.renene.2015.11.065. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S096014811530481X (visited on 12/05/2021).

[61] B. Manobel, F. Sehnke, J. A. Lazzús, I. Salfate, M. Felder, and S. Montecinos, "Wind turbine power curve modeling based on Gaussian Processes and Artificial Neural Networks," en, *Renewable Energy*, vol. 125, pp. 1015–1020, Sep. 2018, ISSN: 09601481. DOI: 10.1016/j.renene.2018.02.081. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148118302258 (visited on 12/05/2021).

[62] R. K. Pandit, D. Infield, and A. Kolios, "Comparison of advanced non-parametric models for wind turbine power curves," en, *IET Renewable Power Generation*, vol. 13, no. 9, pp. 1503–1510, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rpg.2018.5728, ISSN: 1752-1424. DOI: 10.1049/iet-rpg.2018.5728. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2018.5728 (visited on 12/05/2021).

[63] K. Leahy, R. L. Hu, I. C. Konstantakopoulos, C. J. Spanos, and A. M. Agogino, "Diagnosing wind turbine faults using machine learning techniques applied to operational data," in *2016 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Jun. 2016, pp. 1–8. DOI: 10.1109/ICPHM.2016.7542860.

[64] P. Cambron, R. Lepvrier, C. Masson, A. Tahan, and F. Pelletier, "Power curve monitoring using weighted moving average control charts," en, *Renewable Energy*, vol. 94, pp. 126–135, Aug. 2016, ISSN: 09601481. DOI: 10.1016/j.renene.2016.03.031. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148116302154 (visited on 12/05/2021).

[65] B. Jing, Z. Qian, H. Zareipour, Y. Pei, and A. Wang, “Wind Turbine Power Curve Modelling with Logistic Functions Based on Quantile Regression,” en, *Applied Sciences*, vol. 11, no. 7, p. 3048, Jan. 2021, Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/app11073048. [Online]. Available: https://www.mdpi.com/2076-3417/11/7/3048 (visited on 12/05/2021).

[66] D. Astolfi, F. Castellani, and L. Terzi, “Fault prevention and diagnosis through scada temperature data analysis of an onshore wind farm,” *Diagnostyka*, vol. 15, no. 2, pp. 71–78, 2014. [Online]. Available: http://www.diagnostyka.net.pl/Fault-prevention-and-diagnosis-through-SCADA-temperature-data-analysis-of-an-onshore,109967,0,2.html.

[67] Y. Li and Z. Wu, “A condition monitoring approach of multi-turbine based on VAR model at farm level,” en, *Renewable Energy*, vol. 166, pp. 66–80, Apr. 2020, ISSN: 09601481. DOI: 10.1016/j.renene.2020.11.106. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148120318589 (visited on 12/06/2021).

[68] Y. Feng, Y. Qiu, C. J. Crabtree, H. Long, and P. J. Tavner, “Monitoring wind turbine gearboxes,” en, *Wind Energy*, vol. 16, no. 5, pp. 728–740, 2013, ISSN: 1099-1824. DOI: 10.1002/we.1521. [Online]. Available: http://onlinelibrary.wiley.com/doi/abs/10.1002/we.1521 (visited on 12/06/2021).

[69] Z.-Y. Zhang and K.-S. Wang, “Wind turbine fault detection based on SCADA data analysis using ANN,” en, *Advances in Manufacturing*, vol. 2, no. 1, pp. 70–78, Mar. 2014, ISSN: 2095-3127, 2195-3597. DOI: 10.1007/s40436-014-0061-6. [Online]. Available: http://link.springer.com/10.1007/s40436-014-0061-6 (visited on 12/06/2021).

[70] P. Bangalore and L. B. Tjernberg, “An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings,” *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 980–987, Mar. 2015, Conference Name: IEEE Transactions on Smart Grid, ISSN: 1949-3061. DOI: 10.1109/TSG.2014.2386305.

[71] J. Fu, J. Chu, P. Guo, and Z. Chen, “Condition Monitoring of Wind Turbine Gearbox Bearing Based on Deep Learning Model,” *IEEE Access*, vol. 7, pp. 57 078–57 087, 2019, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2912621.

[72] R. Orozco, S. Sheng, and C. Phillips, “Diagnostic Models for Wind Turbine Gearbox Components Using SCADA Time Series Data,” in *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Jun. 2018, pp. 1–9. DOI: 10.1109/ICPHM.2018.8448545.

[73] P. Marti-Puig, A. Blanco-M., M. Serra-Serra, and J. Solé-Casals, “Wind Turbine Prognosis Models Based on SCADA Data and Extreme Learning Machines,” en, *Applied Sciences*, vol. 11, no. 2, p. 590, Jan. 2021, ISSN: 2076-3417. DOI: 10.3390/app11020590. [Online]. Available: https://www.mdpi.com/2076-3417/11/2/590 (visited on 12/06/2021).

[74] C. Dao, B. Kazemtabrizi, and C. Crabtree, “Wind turbine reliability data review and impacts on levelised cost of energy,” en, *Wind Energy*, vol. 22, no. 12, pp. 1848–1871, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.2404, ISSN: 1099-1824. DOI: https://doi.org/10.1002/we.2404. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/we.2404 (visited on 03/05/2021).

[75] M. Du, S. Ma, and Q. He, “A SCADA data based anomaly detection method for wind turbines,” in *2016 China International Conference on Electricity Distribution (CICED)*, ISSN: 2161-749X, Aug. 2016, pp. 1–6. DOI: 10.1109/CICED.2016.7576060.

[76] A. Blanco-M., K. Gibert, P. Marti-Puig, J. Cusidó, and J. Solé-Casals, “Identifying Health Status of Wind Turbines by Using Self Organizing Maps and Interpretation-Oriented Post-Processing Tools,” en, *Energies*, vol. 11, no. 4, p. 723, Mar. 2018, ISSN: 1996-1073. DOI: 10.3390/en11040723. [Online]. Available: http://www.mdpi.com/1996-1073/11/4/723 (visited on 12/06/2021).

[77] A. Purarjomandlangrudi, A. H. Ghapanchi, and M. Esmalifalak, “A data mining approach for fault diagnosis: An application of anomaly detection algorithm,” en, *Measurement*, vol. 55, pp. 343–352, Sep. 2014, ISSN: 02632241. DOI: 10.1016/j.measurement.2014.05.029. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0263224114002504 (visited on 12/06/2021).

[78] C. McKinnon, J. Carroll, A. McDonald, S. Koukoura, D. Infield, and C. Soraghan, “Comparison of New Anomaly Detection Technique for Wind Turbine Condition Monitoring Using Gearbox SCADA Data,” en, *Energies*, vol. 13, no. 19, p. 5152, Oct. 2020, ISSN: 1996-1073. DOI: 10.3390/en13195152. [Online]. Available: https://www.mdpi.com/1996-1073/13/19/5152 (visited on 12/06/2021).

[79] H. Zhao, H. Liu, W. Hu, and X. Yan, “Anomaly detection and fault analysis of wind turbine components based on deep learning network,” en, *Renewable Energy*, vol. 127, pp. 825–834, Nov. 2018, ISSN: 09601481. DOI: 10.1016/j.renene.2018.05.024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148118305457 (visited on 12/06/2021).

[80] W. Yang, C. Liu, and D. Jiang, “An unsupervised spatiotemporal graphical modeling approach for wind turbine condition monitoring,” en, *Renewable Energy*, vol. 127, pp. 230–241, Nov. 2018, ISSN: 09601481. DOI: 10.1016/j.renene.2018.04.059. [Online]. Available:

https://linkinghub.elsevier.com/retrieve/pii/S0960148118304671 (visited on 12/06/2021).

[81] P. Chen, Y. Li, K. Wang, M. J. Zuo, P. S. Heyns, and S. Baggeröhr, “A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks,” en, *Measurement*, vol. 167, p. 108 234, Jan. 2021, ISSN: 02632241. DOI: 10.1016/j.measurement.2020.108234. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0263224120307739 (visited on 12/06/2021).

[82] Á. Encalada-Dávila, C. Tutivén, B. Puruncajas, and Y. Vidal, “Wind Turbine Multi-Fault Detection based on SCADA Data via an AutoEncoder,” en, *Renewable Energy and Power Quality Journal*, vol. 19, pp. 487–492, Sep. 2021, ISSN: 2172038X. DOI: 10.24084/repqj19.325. [Online]. Available: https://www.icrepq.com/icrepq21/325-21-tutiven.pdf (visited on 12/06/2021).

[83] K. Leahy, R. Lily Hu, I. C. Konstantakopoulos, C. J. Spanos, A. M. Agogino, and D. T. J. O’Sullivan, “Diagnosing and PredictingWind Turbine Faults from SCADA Data Using Support Vector Machines,” en, *International Journal of Prognostics and Health Management*, vol. 9, no. 1, Nov. 2020, ISSN: 2153-2648, 2153-2648. DOI: 10.36001/ijphm.2018.v9i1.2692. [Online]. Available: https://papers.phmsociety.org/index.php/ijphm/article/view/2692 (visited on 12/06/2021).

[84] S. Koukoura, J. Carroll, A. McDonald, and S. Weiss, “Comparison of wind turbine gearbox vibration analysis algorithms based on feature extraction and classification,” *IET Renewable Power Generation*, vol. 13, no. 14, pp. 2549–2557, 2019. DOI: https://doi.org/10.1049/iet-rpg.2018.5313. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rpg.2018.5313. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2018.5313.

[85] P. Marti-Puig, A. Blanco-M, J. J. Cárdenas, J. Cusidó, and J. Solé-Casals, “Effects of the preprocessing algorithms in fault diagnosis of wind turbines,” en, *Environmental Modelling & Software*, vol. 110, pp. 119–128, Dec. 2018, ISSN: 13648152. DOI: 10.1016/j.envsoft.2018.05.002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815217302104 (visited on 12/07/2021).

[86] P. Marti-Puig, A. Blanco-M, J. Cárdenas, J. Cusidó, and J. Solé-Casals, “Feature Selection Algorithms for Wind Turbine Failure Prediction,” en, *Energies*, vol. 12, no. 3, p. 453, Jan. 2019, ISSN: 1996-1073. DOI: 10.3390/en12030453. [Online]. Available: http://www.mdpi.com/1996-1073/12/3/453 (visited on 12/07/2021).

[87] Y. Kong, T. Wang, Z. Feng, and F. Chu, "Discriminative dictionary learning based sparse representation classification for intelligent fault identification of planet bearings in wind turbine," en, *Renewable Energy*, vol. 152, pp. 754–769, Jun. 2020, ISSN: 09601481. DOI: 10.1016/j.renene.2020.01.093. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148120301154 (visited on 12/06/2021).

[88] H. Yi, Q. Jiang, X. Yan, and B. Wang, "Imbalanced Classification Based on Minority Clustering Synthetic Minority Oversampling Technique With Wind Turbine Fault Detection Application," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 5867–5875, Sep. 2021, Conference Name: IEEE Transactions on Industrial Informatics, ISSN: 1941-0050. DOI: 10.1109/TII.2020.3046566.

[89] A. Turnbull, J. Carroll, A. McDonald, and S. Koukoura, "Prediction of wind turbine generator failure using two-stage cluster-classification methodology," DOI: 10.1002/we.2391. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/we.2391 (visited on 12/06/2021).

[90] V. Torra, "Trends in Information fusion in Data Mining," en, in *Information Fusion in Data Mining*, J. Kacprzyk and V. Torra, Eds., vol. 123, Series Title: Studies in Fuzziness and Soft Computing, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 1–6, ISBN: 978-3-642-05628-4. DOI: 10.1007/978-3-540-36519-8_1. [Online]. Available: http://link.springer.com/10.1007/978-3-540-36519-8_1 (visited on 12/07/2021).

[91] S. Sheng, "Report on wind turbine subsystem reliability - a survey of various databases (presentation)," Jul. 2013. [Online]. Available: https://www.osti.gov/biblio/1090149.

[92] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15, ISBN: 978-3-540-45014-6.

[93] R. Polikar, "Ensemble Learning," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds., Boston, MA: Springer US, 2012, pp. 1–34, ISBN: 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7_1. [Online]. Available: https://doi.org/10.1007/978-1-4419-9326-7_1.

[94] J. Vanschoren, "Meta-learning," in *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing, 2019, pp. 35–61, ISBN: 978-3-030-05318-5. DOI: 10.1007/978-3-030-05318-5_2. [Online]. Available: https://doi.org/10.1007/978-3-030-05318-5_2.

[95] W. Yang, P. J. Tavner, C. J. Crabtree, Y. Feng, and Y. Qiu, "Wind turbine condition monitoring: Technical and commercial challenges: Wind turbine condition monitoring: Technical and commercial challenges," en, *Wind Energy*, vol. 17, no. 5, pp. 673–693, May 2014, ISSN:

10954244. DOI: 10.1002/we.1508. [Online]. Available: http://doi.wiley.com/10.1002/we.1508 (visited on 02/25/2021).

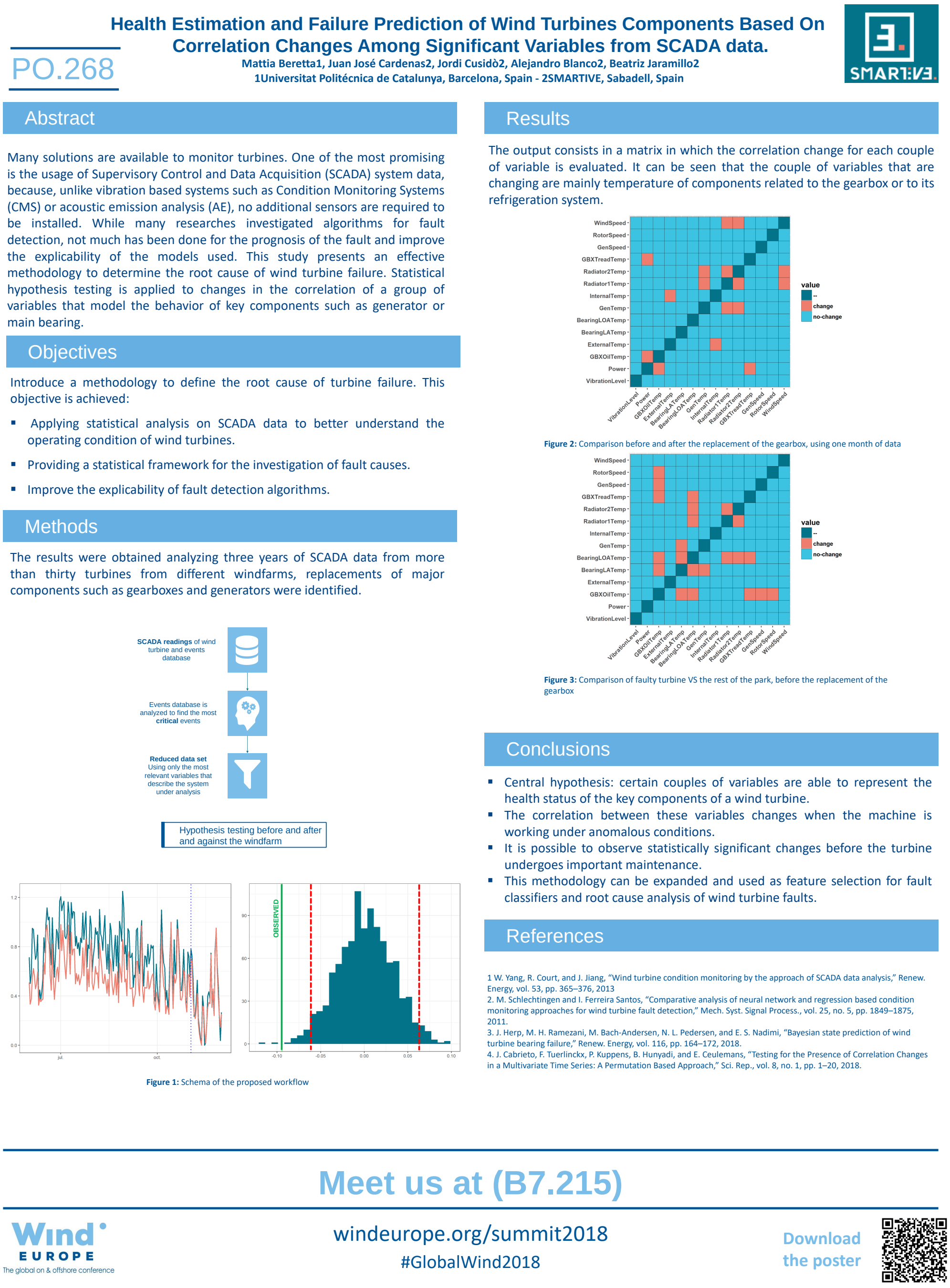
[96] E. Gonzalez, B. Stephen, D. Infield, and J. J. Melero, "On the use of high-frequency SCADA data for improved wind turbine performance monitoring," *Journal of Physics: Conference Series*, vol. 926, p. 012 009, Nov. 2017. DOI: 10.1088/1742-6596/926/1/012009. [Online]. Available: https://doi.org/10.1088/1742-6596/926/1/012009.

[97] E. J. Alvarez and A. P. Ribaric, "An improved-accuracy method for fatigue load analysis of wind turbine gearbox based on SCADA," en, *Renewable Energy*, vol. 115, pp. 391–399, Jan. 2018, ISSN: 09601481. DOI: 10.1016/j.renene.2017.08.040. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148117307851 (visited on 02/18/2021).

A. Appendix

- **Appendix A.1:** Poster WindEurope 2018, Hamburg
- **Appendix A.2:** Posters WindEurope 2019, Bilbao
- **Appendix A.3:** Abstract of the presentation held at WESC 2021, Hannover

A.1 Poster WindEurope 2018, Hamburg



PO.268

Health Estimation and Failure Prediction of Wind Turbines Components Based On Correlation Changes Among Significant Variables from SCADA data.

Mattia Beretta1, Juan José Cardenas2, Jordi Cusidò2, Alejandro Blanco2, Beatriz Jaramillo2
1Universitat Politécnica de Catalunya, Barcelona, Spain - 2SMARTIVE, Sabadell, Spain

Abstract

Many solutions are available to monitor turbines. One of the most promising is the usage of Supervisory Control and Data Acquisition (SCADA) system data, because, unlike vibration based systems such as Condition Monitoring Systems (CMS) or acoustic emission analysis (AE), no additional sensors are required to be installed. While many researches investigated algorithms for fault detection, not much has been done for the prognosis of the fault and improve the explicability of the models used. This study presents an effective methodology to determine the root cause of wind turbine failure. Statistical hypothesis testing is applied to changes in the correlation of a group of variables that model the behavior of key components such as generator or main bearing.

Objectives

Introduce a methodology to define the root cause of turbine failure. This objective is achieved:

- Applying statistical analysis on SCADA data to better understand the operating condition of wind turbines.
- Providing a statistical framework for the investigation of fault causes.
- Improve the explicability of fault detection algorithms.

Methods

The results were obtained analyzing three years of SCADA data from more than thirty turbines from different windfarms, replacements of major components such as gearboxes and generators were identified.

SCADA readings of wind turbine and events database

Events database is analyzed to find the most **critical** events

Reduced data set
Using only the most relevant variables that describe the system under analysis

Hypothesis testing before and after and against the windfarm

Figure 1: Schema of the proposed workflow

Results

The output consists in a matrix in which the correlation change for each couple of variable is evaluated. It can be seen that the couple of variables that are changing are mainly temperature of components related to the gearbox or to its refrigeration system.

Figure 2: Comparison before and after the replacement of the gearbox, using one month of data

Figure 3: Comparison of faulty turbine VS the rest of the park, before the replacement of the gearbox

Conclusions

- Central hypothesis: certain couples of variables are able to represent the health status of the key components of a wind turbine.
- The correlation between these variables changes when the machine is working under anomalous conditions.
- It is possible to observe statistically significant changes before the turbine undergoes important maintenance.
- This methodology can be expanded and used as feature selection for fault classifiers and root cause analysis of wind turbine faults.

References

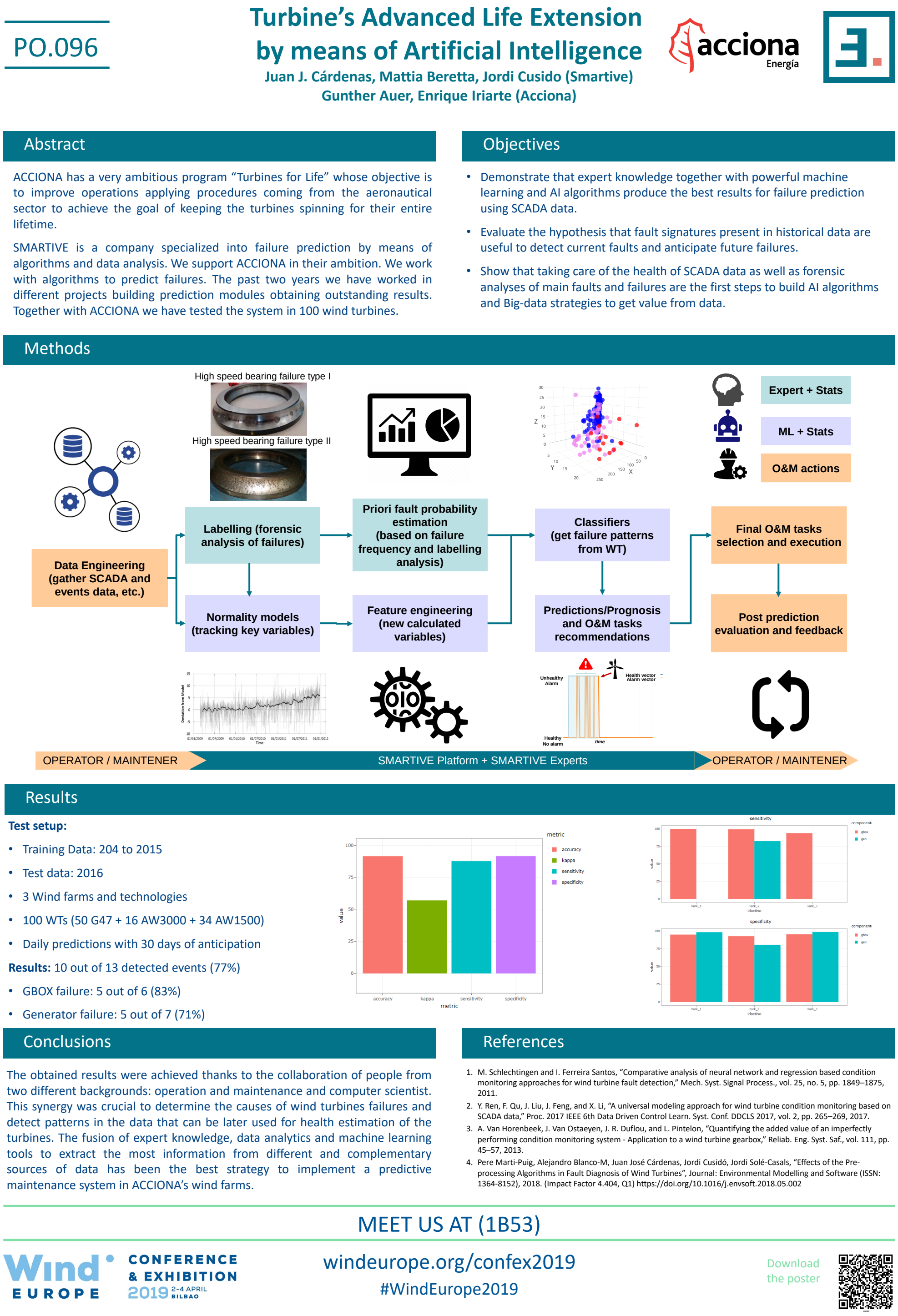
1 W. Yang, R. Court, and J. Jiang, “Wind turbine condition monitoring by the approach of SCADA data analysis,” Renew. Energy, vol. 53, pp. 365–376, 2013
2. M. Schlechtingen and I. Ferreira Santos, “Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection,” Mech. Syst. Signal Process., vol. 25, no. 5, pp. 1849–1875, 2011.
3. J. Herp, M. H. Ramezani, M. Bach-Andersen, N. L. Pedersen, and E. S. Nadimi, “Bayesian state prediction of wind turbine bearing failure,” Renew. Energy, vol. 116, pp. 164–172, 2018.
4. J. Cabrieto, F. Tuerlinckx, P. Kuppens, B. Hunyadi, and E. Ceulemans, “Testing for the Presence of Correlation Changes in a Multivariate Time Series: A Permutation Based Approach,” Sci. Rep., vol. 8, no. 1, pp. 1–20, 2018.

Meet us at (B7.215)

windeurope.org/summit2018
#GlobalWind2018

Download the poster

A.2 Posters WindEurope 2019, Bilbao



PO.096

Turbine's Advanced Life Extension by means of Artificial Intelligence

Juan J. Cárdenas, Mattia Beretta, Jordi Cusido (Smartive)
Gunther Auer, Enrique Iriarte (Acciona)

Abstract

ACCIONA has a very ambitious program "Turbines for Life" whose objective is to improve operations applying procedures coming from the aeronautical sector to achieve the goal of keeping the turbines spinning for their entire lifetime.

SMARTIVE is a company specialized into failure prediction by means of algorithms and data analysis. We support ACCIONA in their ambition. We work with algorithms to predict failures. The past two years we have worked in different projects building prediction modules obtaining outstanding results. Together with ACCIONA we have tested the system in 100 wind turbines.

Objectives

- Demonstrate that expert knowledge together with powerful machine learning and AI algorithms produce the best results for failure prediction using SCADA data.
- Evaluate the hypothesis that fault signatures present in historical data are useful to detect current faults and anticipate future failures.
- Show that taking care of the health of SCADA data as well as forensic analyses of main faults and failures are the first steps to build AI algorithms and Big-data strategies to get value from data.

Methods

High speed bearing failure type I

High speed bearing failure type II

Expert + Stats

ML + Stats

O&M actions

Data Engineering (gather SCADA and events data, etc.)

Labelling (forensic analysis of failures)

Normality models (tracking key variables)

Priori fault probability estimation (based on failure frequency and labelling analysis)

Feature engineering (new calculated variables)

Classifiers (get failure patterns from WT)

Predictions/Prognosis and O&M tasks recommendations

Final O&M tasks selection and execution

Post prediction evaluation and feedback

OPERATOR / MAINTENER

SMARTIVE Platform + SMARTIVE Experts

OPERATOR / MAINTENER

Results

Test setup:

- Training Data: 204 to 2015
- Test data: 2016
- 3 Wind farms and technologies
- 100 WTs (50 G47 + 16 AW3000 + 34 AW1500)
- Daily predictions with 30 days of anticipation

Results: 10 out of 13 detected events (77%)

- GBOX failure: 5 out of 6 (83%)
- Generator failure: 5 out of 7 (71%)

Conclusions

The obtained results were achieved thanks to the collaboration of people from two different backgrounds: operation and maintenance and computer scientist. This synergy was crucial to determine the causes of wind turbines failures and detect patterns in the data that can be later used for health estimation of the turbines. The fusion of expert knowledge, data analytics and machine learning tools to extract the most information from different and complementary sources of data has been the best strategy to implement a predictive maintenance system in ACCIONA's wind farms.

References

1. M. Schlechtingen and I. Ferreira Santos, "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection," Mech. Syst. Signal Process., vol. 25, no. 5, pp. 1849–1875, 2011.
2. Y. Ren, F. Qu, J. Liu, J. Feng, and X. Li, "A universal modeling approach for wind turbine condition monitoring based on SCADA data," Proc. 2017 IEEE 6th Data Driven Control Learn. Syst. Conf. DDCLS 2017, vol. 2, pp. 265–269, 2017.
3. A. Van Horenbeek, J. Van Ostaeyen, J. R. Duflou, and L. Pintelon, "Quantifying the added value of an imperfectly performing condition monitoring system - Application to a wind turbine gearbox," Reliab. Eng. Syst. Saf., vol. 111, pp. 45–57, 2013.
4. Pere Marti-Puig, Alejandro Blanco-M, Juan José Cárdenas, Jordi Cusidó, Jordi Solé-Casals, "Effects of the Pre-processing Algorithms in Fault Diagnosis of Wind Turbines", Journal: Environmental Modelling and Software (ISSN: 1364-8152), 2018. (Impact Factor 4.404, Q1) https://doi.org/10.1016/j.envsoft.2018.05.002

MEET US AT (1B53)

Wind EUROPE CONFERENCE & EXHIBITION 2019 2-4 APRIL BILBAO

windeurope.org/confex2019

#WindEurope2019

Download the poster

PO.091

How ensembling can boost your classifier performances

Mattia Beretta[1-2], Juan José Cardenas[2], Jordi Cusidò[1-2], Isaac Justicia[2]
[1]Universitat Politécnica de Catalunya, Barcelona, Spain - [2]SMARTIVE, Sabadell, Spain

Abstract

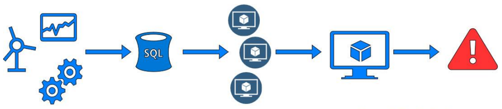
This research investigated the benefits of "**ensembling**" multiple classifiers trying to obtain a better predictive model. This approach is not new in the data analysis field, it has been repeatedly implemented in data-science competition achieving outstanding results. The idea behind ensembling is that **different** algorithms are able to capture various aspects of the same dataset. Aggregating the predictions of various models is possible to **correct biases** and obtain a **more robust predictive structure**. The typical data modeling pipeline includes fitting a large selection of algorithms on the same dataset to evaluate which one obtains the best results. Ensembling is an additional step that takes advantage of the available predictions, stacks them in an additional layer and outputs a class prediction. Different strategies are available for the aggregation of the base algorithms results such as a majority vote, weighted average or even fitting another predictive model such as a logistic regression. All these approaches were investigated in this research to understand the respective pros and cons.

Objectives

- Explore the different option for model ensembling
- Understand the working principle of model stacking and how it can help boosting classifiers performances
- Learn when ensembling should be applied
- Understand which type of models should be included in an ensemble

Methods

The proposed methodology was applied in failure prediction of major components of wind turbines, such as high-speed shafts, gearbox, generators etc. Four years of **SCADA** data from several wind farms has been used. Different machine learning structures such as random forest, support vector machine and gradient boosting machine have been fitted on the available data. The output of the mentioned structures was then fed to a **meta-algorithm** that outputted the final classification label. The evaluation was done on a reserved dataset, not used to train the base learners.



SCADA data is collected by sensors in the turbine and stored in a database

Multiple models are trained on the available data

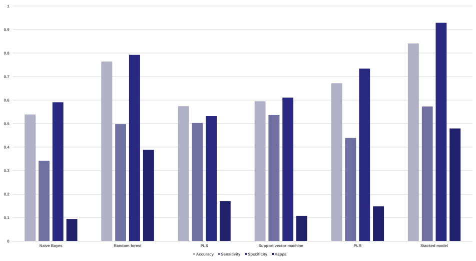
A metamodel is trained on the prediction of the individual models and then used to generate the final prediction

Results

The different ensembling strategies were compared on the same test set, made of a selection of healthy and unhealthy turbines that were not used in the training set.

The tracked classification metrics were Area Under the Curve (AUC), sensitivity, specificity and accuracy. Using an additional algorithm layer is the best strategy to improve classification metrics but it also requires additional computation time and a reserved dataset to fit the second algorithm structure.

Blending different algorithms helps overcoming the limitations of the single algorithms, a careful selection of the utilized models allows to capture various characteristics of the dataset boosting the overall performances.



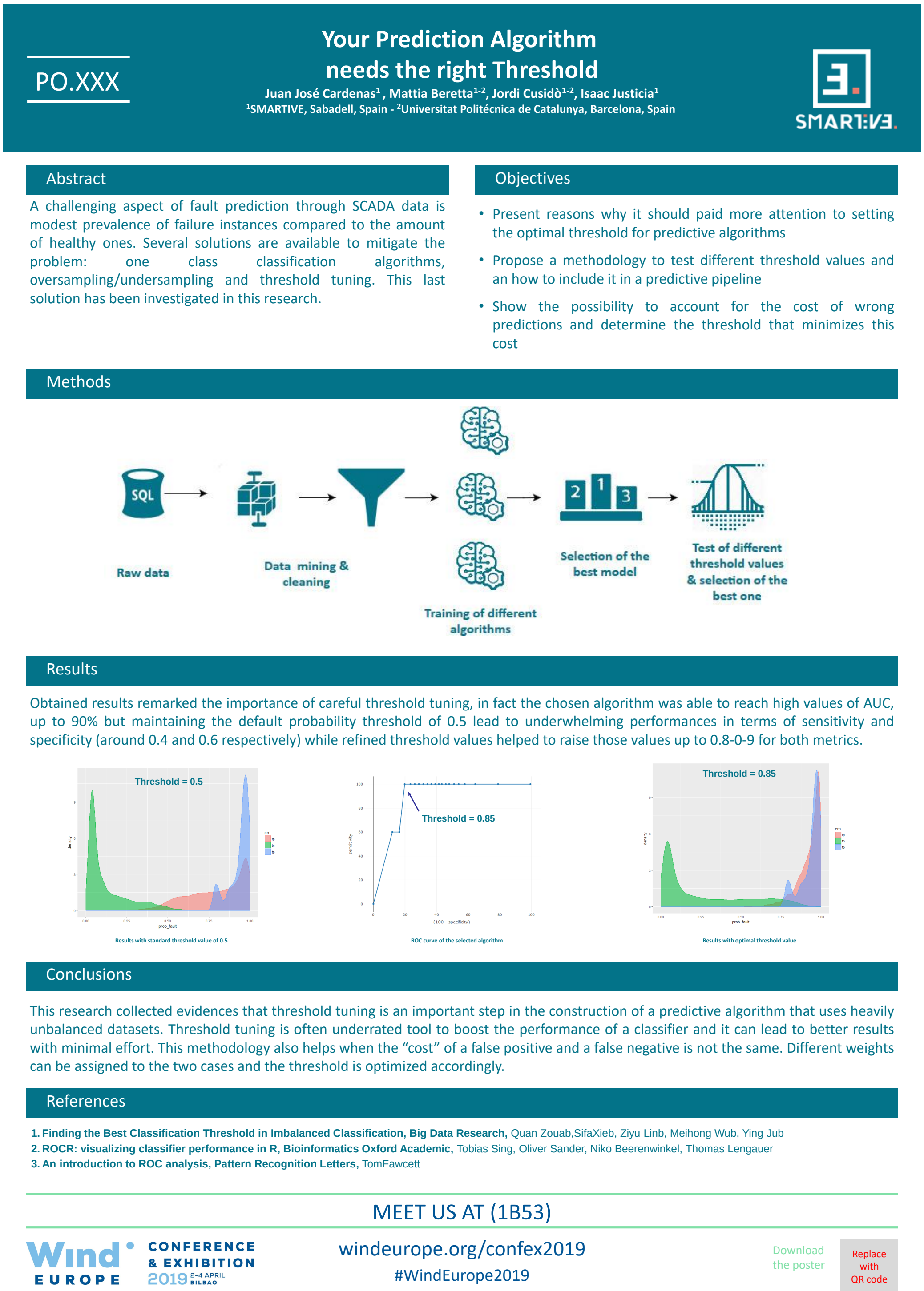
Conclusions

Algorithms ensembling is a valuable strategy to boost the performance of classifiers. It should be considered when a sufficient amount of data is available and the additional training time does not pose a problem. Stacking models is particularly interesting because of the possibility to mix predictions of different predictive structures, specialized in the detection of different fault patterns. This approach also helps reducing the amount of false negatives and false positives, since it is less likely that different predictive structures consistently mis-classify the same input data.

References

1. Issues in stacked generalization, Journal of Artificial Intelligence, Kai Ming Tin and Ian H. Witten
2. Combining Estimates in Regression and Classification, Journal of the American Statistical Association, Michael Leblanc & Robert Tibshirani
3. Is Combining Classifiers with Stacking Better than Selecting the Best One?, Machine Learning, Saso DžeroskiBernard Ženko

MEET US AT (1B53)



PO.XXX

Your Prediction Algorithm needs the right Threshold

Juan José Cardenas[1] , Mattia Beretta[1-2], Jordi Cusidò[1-2], Isaac Justicia[1]

[1]SMARTIVE, Sabadell, Spain - [2]Universitat Politécnica de Catalunya, Barcelona, Spain

SMART:V3.

Abstract

A challenging aspect of fault prediction through SCADA data is modest prevalence of failure instances compared to the amount of healthy ones. Several solutions are available to mitigate the problem: one class classification algorithms, oversampling/undersampling and threshold tuning. This last solution has been investigated in this research.

Objectives

- Present reasons why it should paid more attention to setting the optimal threshold for predictive algorithms
- Propose a methodology to test different threshold values and an how to include it in a predictive pipeline
- Show the possibility to account for the cost of wrong predictions and determine the threshold that minimizes this cost

Methods

Raw data → Data mining & cleaning → → Training of different algorithms → Selection of the best model → Test of different threshold values & selection of the best one

Results

Obtained results remarked the importance of careful threshold tuning, in fact the chosen algorithm was able to reach high values of AUC, up to 90% but maintaining the default probability threshold of 0.5 lead to underwhelming performances in terms of sensitivity and specificity (around 0.4 and 0.6 respectively) while refined threshold values helped to raise those values up to 0.8-0-9 for both metrics.

Results with standard threshold value of 0.5

ROC curve of the selected algorithm

Results with optimal threshold value

Conclusions

This research collected evidences that threshold tuning is an important step in the construction of a predictive algorithm that uses heavily unbalanced datasets. Threshold tuning is often underrated tool to boost the performance of a classifier and it can lead to better results with minimal effort. This methodology also helps when the "cost" of a false positive and a false negative is not the same. Different weights can be assigned to the two cases and the threshold is optimized accordingly.

References

1. **Finding the Best Classification Threshold in Imbalanced Classification, Big Data Research,** Quan Zouab,SifaXieb, Ziyu Linb, Meihong Wub, Ying Jub
2. **ROCR: visualizing classifier performance in R, Bioinformatics Oxford Academic,** Tobias Sing, Oliver Sander, Niko Beerenwinkel, Thomas Lengauer
3. **An introduction to ROC analysis, Pattern Recognition Letters,** TomFawcett

MEET US AT (1B53)

A.3 Abstract WESC Hannover 2021

Quantification of the Information Loss Resulting from Temporal Aggregation of Wind Turbine Operating Data

Mattia Beretta[1,2,], Timo Lichtenstein[3], and Karoline Pelka[3]*

[1] Smartive

[2] Universitat Politécnica de Catalunya

[3] Fraunhofer Institute for Wind Energy Systems IWES

[*] Presenting author

Keywords: operating data | high resolution | aggregation | information loss

In modern wind turbines, a plethora of operating data is acquired with high temporal frequency by a vast number of sensors. However, usually only a selection of these sensor data are stored and typically aggregated as 10-minute average values. Sometimes additionally, the standard deviation is available or/and the maxima and minima measured in these intervals. This kind of storage saves a lot of disk space and bandwidth for data transfer. Unfortunately, much of the information on short timescales is lost, that might otherwise be valuable to better model and track the behavior and condition of the turbine. Using high-frequency operational data constitutes a very promising approach to gain further insights about the wind energy system. Pointing out crucial sensors can therefore help to access the full potential within these high-frequency data. The main objective of our work is to explore and exploit the information contained in high frequency operating data. The goal of this contribution is to quantify the information loss resulting from temporal aggregation of operating data to give recommendations to turbine operators up to which coarseness of resolution only a fraction of information is lost for a certain sensor. Applying these recommendations supports both an optimization of the storage footprint as well as the possibility to carry out several scientific analysis methods. Within this case study, analyses have been carried out on a dataset from a European onshore wind farm containing 12 turbines with a nominal power of 3 MW, for which data of 31 sensors was stored in 1 s resolution. The investigated period covers 15 months of collected data.

In order to assess the effect of aggregating data to a lower temporal resolution, a variable for the information loss has been defined as the difference of an aggregated signal to the original values L(t)agg = s(t)agg − s(t)1s, where s(t) is the signal value of the aggregated or original signal in 1 s resolution, respectively, and t is a timestamp of the original dataset, see Fig. 1. Its aggregated value for each aggregation timeframe is in our case the mean value. Additionally, for all sensors

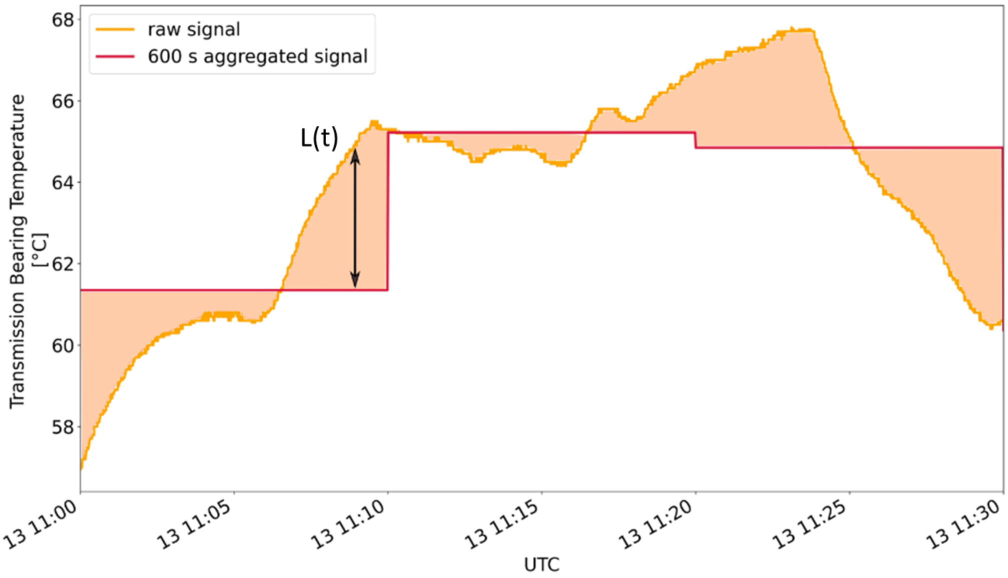


Figure A.1: Illustration of the information loss L for a temporal aggregation of 600 s

the mean information loss per second has been calculated. Calculations have been carried out for multiple levels of temporal aggregation of the data, varying from 5 s to 10 min. To facilitate a comparison of different signals, we have also normalized the data to the interquartile distance of each signal.

The results show only a small loss of information for temperature signals, except the ones related to the gearbox where no more than 25% of the samples had an information loss lower than a tenth of the interquartile distance. The signals with a high information loss are, amongst others, the wind speed and its direction, with more than 50% of the data differing more than a tenth of the interquartile distance from the raw signal.