Utilizing SCADA-Log Data to Improve Normal Behavior Models for Wind Turbine Condition Monitoring

Vorgelegt von
Mohamed Samy Elsisi
aus Ägypten.

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Master of Science
- M.Sc. -
genehmigte Abschlussarbeit.

Gutachter : Prof. Dr. Klus-Robert Müller
Prof. Dr. Thomas Wiegand
Betreuer : M.Sc. Simon Letzgus

Eidesstattliche Versicherung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 01. Mai 2023 .....

Mohamed Samy Elsisi

Abstract

English version of the German “Zusammenfassung”.

Zusammenfassung

Deutsche Version des Englischen “Abstracts”.

Contents

Chapter 1

Introduction

Contents

1.1 Background

In 2020, renewable energy represented 22.1% of energy consumed in the EU [European Commission 2023a]. This percentage is expected to increase drastically in the upcoming years with the target, set by the European Commission, of at least 32% by the year 2030 [European Commission 2023b]. With the increasing number of renewable energy assets being deployed every year, automated condition monitoring solutions are needed for operators to be able to scale up. This need gets more relevant in the case of operating offshore wind farms, where the cost of maintenance relative to the levelized cost of energy (LCOE) is significantly higher compared with onshore [Tautz-Weinert 2017]. Several approaches for condition monitoring were developed in the recent years that use SCADA1 data given its low cost (normally requiring no additional sensors). One of the methods used for condition monitoring using SCADA data is Normal Behavior Modelling (NBM). NBM uses the idea of detecting anomalies from normal operation by empirically modelling a measured parameter, used to reflect the condition of a specific part of the turbine, based on a training phase (usually during a healthy state of the turbine). During operation, the difference between the measured and the modelled/predicted signal is used as indicator for a possible fault. A difference of 0, with some tolerance, reflects normal conditions, whereas a difference greater or less than 0 reflects changed conditions or failures [Tautz-Weinert 2017].

1.2 Motivation

While NBMs using SCADA data were proven capable of predicting failures [Tautz-Weinert 2017], they are treated as black box by the operators since they don't provide any insights regarding the root cause of the failure. Incorporating SCADA log data2 to NBM could help tackle this problem by providing some insights to an anomaly detected by the model in case a relevant warning or failure

message was logged by the SCADA system around the same time. It was also shown that incorporating SCADA logs containing information about operation conditions or control events could help improve the accuracy of the model in case of events unexplainable by the input signals [Letzgus 2020].//

Logs were "Never" treated as input feature in NBMs

Chapter 2

State of the art

Contents

NBM// What was done in the topic of logs?

2.1 Standalone Chapter or better as part of the introduction?

Chapter 3

Methods

Contents

3.1 Dataset

In this section, we will describe the dataset used in this work to train, test and validate the models.

We used open-source data published on the [EDP 2018] *OpenData* web platform and made available for research purposes. The data was collected from the SCADA systems of five different Vestas wind turbines (Turbine 01, 06, 07, 09 and 11) in the same wind park between the years 2016 and 2017 and is made up of the following four subsets: *Signals, Logs, Failures and Metmast*. We will, however, only describe three sets since *Metmast* was not used in this work.

3.1.1 Signals

The *Signals* dataset contains 10-min-averaged data collected from the wind turbines' sensors installed at the major components (e.g., gearbox, generator, transformer,..) and power meters to measure temperatures, angles, wind and rotational speeds, power production,... This dataset was the most crucial for this work since it provides

Type of signal	Signals
Average temperature (°C)	Generator, Generator bearings, Hydraulic group oil, Gearbox oil, Gearbox bearing on the high-speed shaft, Nacelle, HV transformer, Ambient temperature,..
Average production value	Active power (Wh), Reactive power (VArh), Power according to grid (kW),..

Table 3.1: Sample signals found in the Signals dataset

Type of log event	Sample log event
Alarm log	*"High temperature brake disc"* *"High pres offlin:____RPM/ ___°C"*
Warning log	*"Yaw Position is changed: __°"* *"Low Battery Nacelle"*
Operation and System log	*"External power ref.:____kW"* *"GearoilCooler _, gear: ___°C"*

Table 3.2: Sample log events found in the Logs dataset

information that reflects the status of the turbine operation which is needed to perform automated condition monitoring and predictive maintenance.
Table 3.1 shows a sample of the 81 signals included in this dataset.

3.1.2 Logs

In this dataset, events logged by the SCADA system are collected in non-fixed intervals. The events recorded by the system are divided into three categories: Alarm log, Warning log and Operation and System log. According to the VestasOnline Enterprise user manual [Vestas 2016], alarms are system notifications that alert operators to an error scenario that has forced a wind turbine to cease normal operation and transition to one of three operational states: Pause, Stop, or Emergency (one of the following three acknowledgments is needed to resume operation: Local acknowledgment from the controller unit of the turbine, Remote acknowledgment from VestasOnline®, or Automatic acknowledgment), whereas warnings are system messages that indicate an irregularity that requires attention but does not cause the turbine to immediately cease normal operation and exit the Run state.

3.1.3 Failures

The Failures dataset contains the history of failures, inspections, or maintenance that occurred in the turbines and was manually recorded by technicians. Each record reports the time of the event, component (e.g., Generator, Hydraulic group,..), and a text description of the failure or event (e.g., "Generator replaced", "Oil leakage in Hub",..).

This dataset was used in backtesting to validate the models' capability of detecting failures early.

3.2 Machine learning models

In this section, we will demonstrate the architecture of the machine learning models used in our experiments.

3.2.1 Linear regression

Sir Francis Galton proposed the idea of linear regression in 1894 [Galton 1894]. Linear regression is used for analyzing the linear relationship between one or more independent variables (X) and a dependent variable (Y). The dependent variable Y must be continuous, whereas the independent variables can be continuous or categorical.
When the relationship between the dependent variable and the independent variables is assumed to be linear and there are a small number of independent variables, linear regression is usually used. Linear regression is easy to use and understand, and it can be used to make predictions or find relationships between variables.
In the example of normal behavior modeling for a wind turbine component, the dependent variable can be defined as the component's temperature and the independent variables as a set of weather and turbine conditions measures (e.g., wind speed, ambient temperature, production value, other components' temperatures,..) that have either a direct or indirect effect on the target component.

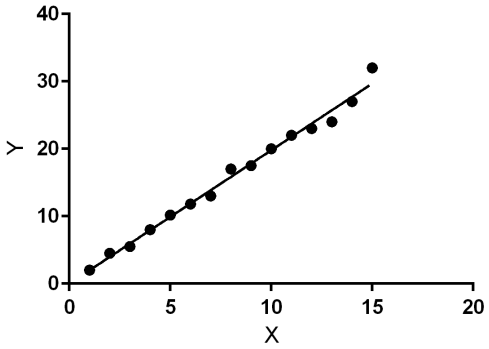

Figure 3.1: Example linear regression

The way the independent variables are chosen is usually done by measuring the correlation coefficients between available features in a dataset and the target feature and then selecting the features having a high correlation coefficient. Depending on the problem setting, other features can be also considered based on domain knowledge, especially when dealing with a mechanical system as in the case of this work.

A good example of this would be the incorporation of the ambient temperature measurement as an input feature—even if it does not highly correlate with the target feature—to make sure that your model generalizes, when trying to predict a component's temperature throughout the year, by considering the effect of seasonality (temperatures are expected to be higher in summer than in winter).
In this work, we selected input features based on both domain knowledge and correlation coefficients. We used Kendall's method to measure the rank correlation [Kendall 1938]. In contrast to Pearson's correlation coefficient, Kendall's rank correlation can capture both linear and non-linear dependency between two variables by measuring the monotonic relationship. In addition to that, variables don't have to be normally distributed when using Kendall's method.

3.2.2 Deep learning

Although multiple linear regression models are capable of fitting the data with high accuracy in many applications (e.g., [Wang 2019]), they are, by definition, not capable of capturing more complex non-linear dependencies. In addition to that, linear regression may not be appropriate when there are a significant number of independent variables. Deep learning may be a better approach in these situations. After obtaining better results with it (see 4.1), we decided to train the normal behavior models on a feed-forward neural network (for a comprehensive review of deep learning and neural networks, see [Schmidhuber 2014]) having the architecture shown in fig. 3.2.

3.3 Log analysis

In this section, we will describe the different approaches we propose to utilize SCADA log messages and incorporate them into normal behavior models. In summary, we introduce three different ways for utilizing SCADA log messages: Extracting input features for normal behavior models, Data filtering, and Visualization of warnings. We will explain each approach in depth.

3.3.1 Extracting input features

Most machine-learning architectures can only work with vector-shaped numerical inputs. Given that there are limited resources in the research field on how to generate numerical vectors from wind turbine SCADA system logs (see 2), we came up with two methods that were proven capable of not only generating input features for machine-learning normal behavior models but also improving their accuracy (see 4): 1. our Novel method based on domain knowledge and 2. Utilizing an open-source framework for analyzing log data called LogPAI. We will discuss each method in detail.

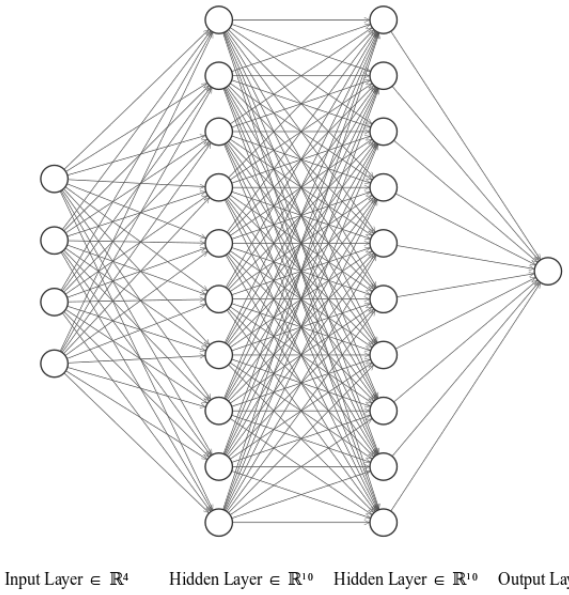

Figure 3.2: Architecture of normal behavior neural network model used in this work. *P.S.: The input layer shape will vary based on the experiment and the number of input features.*

3.3.1.1 Novel method

Background:
We scanned through the different log messages available in the dataset looking for information that reflects the turbine state and might help the normal behavior model fit the data more accurately. Since normal behavior models monitor the state of a component by monitoring its temperature, we narrowed the search down to operation and system logs that reflect events causing a change of temperature in major components. We, then, ended up with a category of logs that shows the states of internal or external ventilators of some components (see 3.3). Being parts of the cooling systems of major components, fans or ventilators must affect the component's temperature.

Log text template	Log text sample
Gen. ext. vent. _, temp:___°C	Gen. ext. vent. 2, temp:65°C
Gen. int. vent. _, temp:___°C	Gen. int. vent. 1, temp:50°C
HV Trafo. vent. _, temp:___°C	HV Trafo. vent. 0, temp:2°C
Nac.vent._, nac/gear:___/___°C	Nac.vent.3, nac/gear:43/ 54°C

Table 3.3: Logs

Indeed, our analysis showed a clear relationship between the state of a ventilator and the temperature of its turbine component. As shown in 3.3, at low temperatures

of the generator bearings, the internal ventilator will switch off. The bearings will then heat up which, in turn, causes the ventilator to turn on which cools the bearings down, and so on...

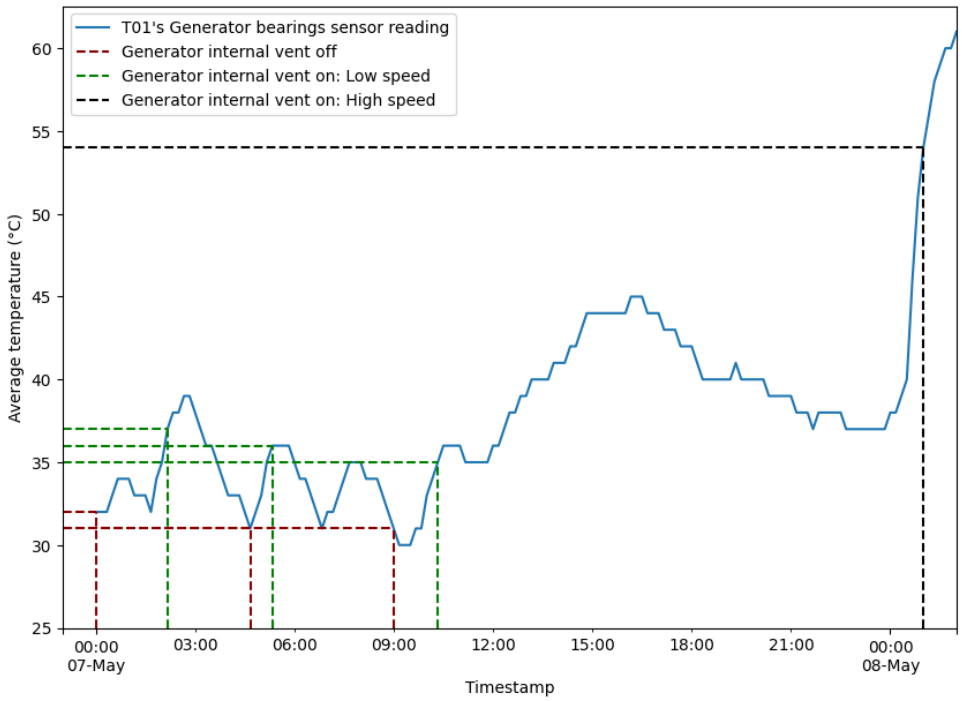

Figure 3.3: Generator internal vent control signals and their effect on the generator bearings temperature

Method:
Analyzing the log texts of interest (e.g., *Gen. ext. vent. 2, temp:65°C*), we deduce that they provide three pieces of information: 1. Description of the ventilator (e.g., *Gen. ext. vent.*), 2. State of the ventilator (*0, 1, 2 or 3*), 3. Temperature of the turbine component the ventilator is installed in (e.g., *65°C*).

Since the component temperature is regularly provided as a SCADA signal, we decided to focus on the other two parts of the log messages. Our method simply filters log messages containing the word "vent." and creates a new feature for every ventilator (1.) found in the data having its state (2.) as a value.
In contrast to the signals data fixed rate of occurrence (10 min), the generated log feature has an inconsistent frequency (the SCADA system creates a new log entry only when a ventilator changes states). We join both datasets by taking the value of the last occurrence in the log feature vector within a 10-min window relative to a signal reading. Gaps in the log feature columns in the resulting dataset are then filled by propagating the last valid observation forward to the next valid (a ventilator has the same state as long as it hasn't changed).

3.3.1.2 Utilizing LogPAI

LogPAI (Log Analysis and Intelligence) is a study project and open-source platform for analyzing and managing log data [Lyu 2019]. Tsinghua University researchers started the project, which focuses on developing efficient algorithms and tools for log analysis, anomaly detection, and log data visualization. LogPAI includes a complete suite of log analysis and processing tools such as Logparser, Loglizer, and Logreduce. These applications can assist users in preprocessing and parsing raw log data, detecting anomalies and patterns, and summarising log data concisely and understandably.

TODO: 1. Explain how the LogParser (Drain) works, what the format of the parsed log messages look like and why Drain -> Because online log parser. 2. Explain how LogLizer creates a feature from parsed logs

3.3.2 Data filtering

TODO: Stoppages (log as filter)

3.3.3 Visualization of warnings

TODO: Easy: look for any warning/alarm message (having the word "hot" or "high temperature") related to the target component (having the name of the component in the message) and append to the anomaly on the operation dashboard

3.4 Anomaly detection

TODO: Define anomalies (what is an anomaly in general?).
Discuss anomaly detection methods used in other papers
Argue why there's no standard way of defining/measuring an anomaly
Describe the way our anomaly detection works.

3.4.1 Alarms

TODO: Discuss the difference between Anomaly and alarm and why we wanna limit the number of alarms being sent to operators (false alarms are costly!)

3.5 Summary

TODO: PUSH TO THE TOP// Diagram of all methods put together: ML model + log feature + Anomaly detection + Alarms,...

Chapter 4

Experiments

Contents

4.1 Experiment I: Benchmark

4.1.1 Research question

The aim of this experiment is to test the ability of NBM, as defined in the literature, to early-detect failures of a faulty turbine using the ERP dataset. The final model architecture that we pick after running this experiment will serve as a baseline model used as a benchmark in the later experiments.

4.1.2 Setup

The following elements were used in this experiment:

- **Machine learning models:** Linear regression (baseline) and feed-forward neural network
- **Target wind turbine:** T09
- **Dataset:** Training/healthy period: 01/01/2016 - 15/02/2016, Testing/faulty period: 16/02/2016 - 18/10/2016
- **Input features:** Nac_Temp_Avg, Amb_Temp_Avg, Gen_RPM_Avg, Prod_LatestAvg_TotActPwr (or use verbose names of signals: the average temperature in the nacelle, average ambient temperature, average generator rpm, total active power)

- **Target feature:** Gen_Bear_Temp_Avg (Average temperature in generator bearing 1 (Non-Drive End))
- **Recorded failure:** *"Generator bearings replaced on October 17, 2016, 9:19 AM"*
- **Logs used:** None

4.1.3 Results

According to the results documented in 4.2, we conclude that both NBMs are capable of predicting the failure in the monitored part. We will, however, use only the feed-forward network model as a benchmark since it outperformed the linear regression model.

Comparison metric	**Measure for linear regression**	**Measure for feed-forward network**
RMSE		
First-detected anomaly timestamp		
Number of anomalies detected		

Table 4.1: Experiment I results: Metrics used to compare between the benchmark models

4.2 Experiment II: Incorporating log features into NBM applied on healthy turbine

4.2.1 Research question

The aim of this experiment is to quantitatively (RMSE) and qualitatively (number of false alarms) measure the effect of incorporating SCADA-log-based features into the benchmark NBM.

4.2.2 Setup

The following elements were used in this experiment:

- **Machine learning models:** Feed-forward neural network with single target features and Feed-forward neural network with multiple target features
- **Target wind turbine:** T01
- **Dataset:** Training/healthy period: 01/09/2016 - 31/12/2016, Testing period: 01/01/2017 - 31/12/2017

- **Input features (SCADA signals):** Nac_Temp_Avg, Amb_Temp_Avg, Gen_RPM_Avg, Prod_LatestAvg_TotActPwr (or use verbose names of signals: the average temperature in the nacelle, average ambient temperature, average generator rpm, total active power)
- **SCADA-log-based input features:** Operation and System log messages containing the word "vent", which resulted in four different features extracted from the following components: Generator external vent, Generator internal vent, High-voltage transformer vent, and Nacelle vent
- **Target feature for single-output model:** Gen_Bear_Temp_Avg (Average temperature in generator bearing 1 (Non-Drive End))
- **Target features for multiple-output model:** All signals whose names contain the keywords "Gen" and "Temp": 'Gen_Bear_Temp_Avg', 'Gen_Phase1_Temp_Avg', 'Gen_Phase2_Temp_Avg', 'Gen_Phase3_Temp_Avg', 'Gen_SlipRing_Temp_Avg', 'Gen_Bear2_Temp_Avg'
- **Recorded failure:** No generator-related recorded failures (hence the assumption that the turbine is healthy)

4.2.3 Results

According to the results documented in 4.2, we conclude that both NBMs are capable of predicting the failure in the monitored part. We will, however, use only the feed-forward network model as a benchmark since it outperformed the linear regression model.

Comparison metric	**Measure for linear regression**	**Measure for feed-forward network**
RMSE		
First-detected anomaly timestamp		
Number of anomalies detected		

Table 4.2: Experiment I results: Metrics used to compare between the benchmark models

Chapter 5

Conclusions

Appendix A

Appendix I

Bibliography

[EDP 2018] EDP. *EDP Open Data*. https://opendata.edp.com/opendata/en/data.html, July 2018. Accessed: 2023-01-29. (Cited on page 5.)

[European Commission 2023a] European Commission. *Renewable energy statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics, 03 2023. Accessed: 2023-03-26. (Cited on page 1.)

[European Commission 2023b] European Commission. *Renewable energy targets*. https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-targets_en, 2023. Accessed: 2023-03-26. (Cited on page 1.)

[Galton 1894] Sir Galton Francis. Natural inheritance. New York, Macmillan and co, 1894. https://www.biodiversitylibrary.org/bibliography/46339. (Cited on page 7.)

[Kendall 1938] M. G. Kendall. *A NEW MEASURE OF RANK CORRELATION*. Biometrika, vol. 30, no. 1-2, pages 81–93, 06 1938. (Cited on page 8.)

[Letzgus 2020] Simon Letzgus. *Finding meaningful representations of SCADA-log information for data-driven condition monitoring applications*. 12 2020. (Cited on page 2.)

[Lyu 2019] Michael R. Lyu, Jieming Zhu, Pinjia He, Shilin He, Jinyang Liu, Zhuangbin Chen, Yintong Huo, Yuxin Su and Zibin Zheng. *LogPAI*. https://logpai.com/, 2019. (Cited on page 11.)

[Schmidhuber 2014] Jürgen Schmidhuber. *Deep Learning in Neural Networks: An Overview*. CoRR, vol. abs/1404.7828, 2014. (Cited on page 8.)

[Tautz-Weinert 2017] Jannis Tautz-Weinert and Simon Watson. *Using SCADA data for wind turbine condition monitoring - A review*. IET Renewable Power Generation, vol. 11, pages 382–394, 03 2017. (Cited on page 1.)

[Vestas 2016] Vestas. *VestasOnline Enterprise User Guide*. https://voe.vestas.com/public/UserManual.pdf, 02 2016. Accessed: 2023-02-06. (Cited on page 6.)

[Wang 2019] Xiaofeng Wang, Guoliang Lu and Peng Yan. *Multiple regression analysis based approach for condition monitoring of industrial rotating machinery using multi-sensors*. In 2019 Prognostics and System Health Management Conference (PHM-Qingdao), pages 1–5, 2019. (Cited on page 8.)