CARE to Compare: A real-world dataset for anomaly detection in wind turbine data

Christian Gück[a,*], Cyriana M.A. Roelofs[a], Stefan Faulstich[a]

[a]*Fraunhofer IEE, Joseph-Beuys-Straße 8, 34117 Kassel, Germany*

Abstract

Anomaly detection plays a crucial role in the field of predictive maintenance for wind turbines, yet the comparison of different algorithms poses a difficult task because domain specific public datasets are scarce. Many comparisons of different approaches either use benchmarks composed of data from many different domains, inaccessible data or one of the few publicly available datasets which lack detailed information about the faults. Moreover, many publications highlight a couple of case studies where fault detection was successful. With this paper we publish a high quality dataset that contains data from 36 wind turbines across 3 different wind farms as well as the most detailed fault information of any public wind turbine dataset as far as we know. The new dataset contains 89 years worth of real-world operating data of wind turbines, distributed across 44 labeled time frames for anomalies that led up to faults, as well as 51 time series representing normal behavior. Additionally, the quality of training data is ensured by turbine-status-based labels for each data point. Furthermore, we propose a new scoring method, called CARE (Coverage, Accuracy, Reliability and Earliness), which takes advantage of the information depth that is present in the dataset to identify a good all-around anomaly detection model. This score considers the anomaly detection performance, the ability to recognize normal behavior properly and the capability to raise as few false alarms as possible while simultaneously detecting anomalies early.

Keywords: benchmark, anomaly detection, wind turbines, predictive

*Corresponding author

Email addresses: christian.gueck@iee.fraunhofer.de (Christian Gück), cyriana.roelofs@iee.fraunhofer.de (Cyriana M.A. Roelofs), stefan.faulstich@iee.fraunhofer.de (Stefan Faulstich)

maintenance, fault detection, condition monitoring

1. Introduction

Wind energy plays a crucial role in the transition to renewable energy, but monitoring and maintaining wind farms and turbines is a costly challenge. These farms are often located in regions with challenging weather conditions, leading to complex operating conditions and increased risk of unexpected failures and downtime. Over the past decade, various approaches for condition monitoring, many of which focus on early fault detection using Supervisory Control and Data Acquisition (SCADA) data, have been investigated [1–3].

A common method to detect component failures early is anomaly detection (AD), which identifies outliers or other anomalous patterns in the data. In the context of wind turbines (WTs), most AD techniques utilize data from the SCADA system, failure logs, vibration data and occasionally status and maintenance logs [4]. This paper specifically focuses on AD models based on SCADA data which are validated using additional failure information.

While there have been several benchmarks [5, 6], reviews [7] and comparisons [8] of general AD-algorithms, most of them use data from a wide variety of domains like spacecraft, medical applications and IT-related data. However, efforts on wind energy specific AD are usually based on non-public inaccessible data. For example [9, 10] use inaccessible data from wind farms that are located in China, and [11–13] use data from anonymized offshore wind farms. These are 5 recently published examples, which lack the ability for meaningful comparisons between the proposed fault detection algorithms. Also, inaccessible data prevents reproducibility of presented results. Some studies have used public wind energy datasets (for example [4, 14–20]), but they lack comprehensive information about anomalies or component faults. The lack of extensive public datasets with both SCADA time series and failure information is a significant limitation in the field of WT SCADA data analysis. To enable meaningful comparisons between AD algorithms in the wind energy domain, new public benchmark datasets are necessary.

The main contribution of our work is the publication of the most extensive WT SCADA dataset[1] for AD yet. This includes high dimensional data

[1]The data can be found on "Fordatis": `http://dx.doi.org/10.24406/fordatis/343` and on "Zenodo": `https://zenodo.org/doi/10.5281/zenodo.10958774`

from multiple wind farms, information about the WT status at all times, labeled anomalies with annotated starts and ends and additional fault descriptions. Because the data stems from real-world operating wind farms it had to be anonymized with the focus on minimizing the loss of useful information and maximizing the meaningfulness of this dataset for AD and predictive maintenance.

In addition to the dataset we also provide a sophisticated score, the Coverage Accuracy Reliability Earliness (CARE)-score, for evaluating AD-algorithms on this and similar datasets. This score takes into account four key aspects of a high-quality AD model for predictive maintenance. In combination with the dataset this score provides the possibility to compare a variety of different AD-algorithms, from unsupervised to semi-supervised techniques, designed for early fault detection in WT.

The content of this paper divides into the following sections. At first we give an overview about the related work in section 2. After that we introduce the dataset in section 3 by giving information about the layout, the requirements we set for the quality of the data, the labeling process and the anonymization actions that were taken. Following this we provide our scoring idea together with a mini-benchmark of a few selected AD-algorithms in section 4. Finally a summary concludes this work in section 5.

2. Related work

In the field of AD benchmark data, many studies focus on dataset compositions from various different domains and use cases. Many benchmark datasets also include a mix of artificial data and data from real-world applications. While [5, 7, 21, 22] study a wide spectrum of different AD algorithms for a broad collection of data types, there are also several AD benchmarks

---

Nomenclature:

WT wind turbine

AD anomaly detection

NBM normal behaviour model

SCADA Supervisory Control and Data Acquisition

WS weighted score

Acc accuracy-score

CARE Coverage Accuracy Reliability Earliness

NN neural network

AE autoencoder

RE reconstruction error

AUC area under the curve

ROC receiver operating characteristic curve

which focus on time series data. One example of such benchmarks is the widely used and cited Numenta benchmark [6], which provides a collection of datasets. A more recent and more comprehensive evaluation of AD in time series is found in [8], where over 71 algorithms were evaluated on more than 900 time series.

In the scope of AD for WTs the time series datasets mentioned above are usually too broad to be used for evaluation in this specific context. Also, many are either univariate or synthetic time series and therefore not applicable to AD in SCADA-data. Unfortunately, most domain-specific evaluations are conducted on inaccessible data that were provided only for the research in which they are used. As mentioned in the introduction, there are plenty examples of studies which use such datasets.

There are only a handful of open datasets containing WT SCADA-data. The studies [23, 24] give an overview about existing datasets for WTs. Additionally the Git-repository [25] summarizes some currently existing datasets although some of the listed datasets, such as the SCADA-data of the ENGIE wind farm "La Haute Borne", are not available anymore. The most relevant public dataset in the context of AD for early fault detection is provided by the EDP open data platform [26], since it is, as far as the authors know, the only one containing information about WT faults in addition to the SCADA-data. The faults are provided in form of a start timestamp for some turbine faults. This dataset was used in the fault detection challenge "hack the wind" [27] hosted by EDP which is mentioned and evaluated together with the "WeDoWind"-challenge [28] in [20] and [19]. These challenges focus in particular on the evaluation of AD-algorithms based on maintenance cost and potential savings that could be achieved through predictive maintenance. Furthermore, several studies on fault detection have used this data [4, 15, 17, 18]. Although the EDP-dataset is widely used, its level of detail regarding the fault information is small, especially in comparison to the inaccessible datasets mentioned before.

The lack of publicly available SCADA-datasets of WTs is also acknowledged in [3], which highlights that it is a constraint in the progress of WT SCADA applications. Additionally, the absence of publicly available datasets containing real-world anomalies is recognized as a significant obstacle in the development of AD in general, as it may not adequately reflect the performance of methods in real-world applications [7, 22].

But there is not only need for additional domain specific public datasets, the data quality and the level of detail also plays an important role. As

pointed out by [29] many AD benchmark datasets suffer from flaws that limit their significance. The main flaws are defined as the flaw of "Triviality", "Unrealistic Anomaly Density", "Mislabeled Ground Truth" and the "Run-to-Failure Bias". In the case of publicly available wind SCADA datasets, one common issue is the absence of labels, particularly regarding fault information.

Considering the flaws in datasets, scoring for AD algorithms poses a difficult challenge. Many studies utilize standard classification metrics such as accuracy, precision, recall, or the area under the curve (AUC) of the receiver operating characteristic curve (ROC) [5, 11, 13, 30].

While it is possible to evaluate AD algorithms using the AUC-ROC score for all possible thresholds, for most practical applications it is much more useful to have a high F-Score, or a related score. In [31] several variants of F-scores are compared. The standard pointwise F-Score is the simplest, but for most use cases, the interest lies in detecting anomaly events, i.e., a continuous set of anomalous time points, rather than individual time points. A composite F-score is introduced, a modification of the classic F-score that takes into account anomaly events through event-wise recall.

Another approach, presented in [32], modifies the classic AUC-ROC metric by generalizing the concept of the ROC to the Preceding-Window-ROC, thereby adjusting the measure to better fit AD evaluations on time series data from an event-based perspective.

Finally, the Numenta Benchmark [6] defines a score that is supposed to measure the performance of more general AD models for time series data across different domains. The score is based on 5 key-aspects of a good AD model: "detection of all anomalies", "early detection of anomalies", "no false alarms", "uses only real time data" and "automation across all different datasets".

Based on the provided overview of related work, this paper contributes to the progress of AD for predictive maintenance on WTs by introducing a new public dataset that offers more detailed information about turbine faults and associated anomalies. Furthermore, the new dataset addresses the flaws identified by [29], although the potential for mislabeled ground truth cannot be completely eliminated in this context, as the start of anomalous behavior is often unclear. The flaw of triviality is tackled by the inclusion of complex anomalies from real-world WTs based on feedback of the wind

farm operators. Additionally, the proposed CARE-score, which differs from standard classification metrics, draws inspiration from the first three key aspects of the Numenta score and the composite F-Score from [31], while distinct adaptions and further developments have been made to better fit the specific use case of AD for predictive maintenance on WTs.

3. Data

In this section, we describe the new dataset provided with this paper. First, we discuss the requirements for a good dataset for AD in WTs in section 3.1. Then, we provide an overview of the data published in section 3.2, including general statistics such as the number of anomalous events and features, as well as data quality. In section 3.3, we explain the process of labeling each time series and datapoint, and in section 3.4 the anonymization process is described.

3.1. Data requirements

During the process of selecting data for this benchmark dataset, seven requirements were defined to ensure the quality and significance of comparisons of AD algorithms for AD in WTs. The requirements are as follows:

1. The dataset must contain as many anomaly events as possible.
2. The dataset must contain different wind farms.
3. The dataset must contain different fault types.
4. The dataset must be balanced, i.e. contain enough prediction data representing normal behavior.
5. Every sub-dataset must contain enough normal behavior data in the intended training time frame. If at least 2/3 of the training data are normal behavior data we define the sub-dataset to be sufficient.
6. Every sub-dataset must contain at least one whole year worth of data, to be able to learn seasonality-related effects.
7. Every anomaly must have an assigned start timestamp. The anomaly end is the start of a turbine fault.

While requirements 1 to 3 are necessary to test the generalization ability of AD algorithms, requirement 4 enables tests for the ability to learn normal behavior effectively. This is particularly important for the evaluation of

normal behaviour models (NBMs). Additionally, requirements 5 and 6 ensures the quality of the training data, to guarantee an NBM can be trained. Finally, requirement 7 allows for the evaluation of AD models using classification measures. These requirements ensure that the dataset is of high-quality, comprehensive and balanced to train a proper NBM, with detailed labels to validate the model. All these properties are also relevant for the definition of the score introduced in section 4.1.

3.2. Dataset

The data consists of 95 datasets, containing 89 years of SCADA time series distributed across 36 different WTs from the three wind farms A, B and C. The data for Wind farm A is based on the earlier mentioned EDP-data [26], and consists of 5 WTs of an onshore wind farm in Portugal. From this data 22 datasets were selected to be included in this data collection. The other two wind farms are offshore wind farms located in Germany. All three datasets were anonymized as described in section 3.4. The overall dataset is balanced, as 44 out the 95 datasets contain a labeled anomaly event and the other 51 datasets represent normal behavior. Each dataset is provided in form of a csv-file with columns defining the features and rows representing the data points of the time series.

The datasets consist of SCADA time series data for each turbine, with a resolution of 10 minutes. Each dataset includes one year worth of data for training a model, as well as 4 to 98 days of prediction data.

The prediction data is divided into an event time frame, with varying amounts of padding data before and after the event. This padding is used to prevent guessing the event label ("anomaly" or "normal") based on the amount of prediction data.

The number of features in the datasets varies depending on the wind farm. Wind farm A has 86 features, wind farm B has 257 features, and wind farm C has 957 features. In addition to the sensor data features, each time series includes 5 descriptive features: a row ID and a timestamp, an asset ID that identifies the WT, a "train_test" column indicating whether the row belongs to the training or prediction data, and a status-ID indicating the turbine status at the timestamp.

The remaining features represent sensor measurements. For each sensor, the 10-minute average value is available. Some sensors also have additional information in the form of 10-minute minimum, maximum, and standard deviation values. The original sensor names have been replaced in order to

anonymize the data, as described in section 3.4. Only features that describe power, reactive power or wind speed are recognizable by their name. To accommodate for the loss in information, additional descriptions are provided for every sensor. These descriptions include a brief text, the unit of the sensor as well as boolean indicators that imply whether the sensor represents a regular sensor signal, a counter or an angle. The most important statistics of the data are summarized in table 3.2. The rows "Anomaly events" and "Normal behavior" describe the number of datasets containing an anomaly event and without anomalies respectively.

Regarding the data quality there are two challenges. The data for wind farm B and C was provided by the operator with 0-values replacing all missing values, so large amounts of consecutive 0-values must be treated with caution. Secondly, note that the status values for wind farm B and C may be inconsistent; often the status is only logged when it changes, which may fail if there is a brief communication error. Also, the status values for wind farm A were derived based on the EDP fault logbook, which only contained start timestamps of the faults (see section 3.3). It is therefore advisable to check the power and wind speed values in addition to the status values to determine whether the turbine has indeed been operating normally.

	Wind Farm A	Wind Farm B	Wind Farm C	Overall
Turbines	5	9	22	36
Datasets	22	15	58	95
Anomaly events	11	6	27	44
Normal behavior	11	9	31	51
Features	86	257	957	-
Sensors	54	63	238	-

3.3. Data labeling

The data is labeled on two levels. The first level are the so-called event labels. If a dataset contains an anomaly event inside the prediction time frame, the dataset is labeled as an anomaly. If this is not the case it is labeled as normal. The anomaly labels have been determined either based on direct feedback by the wind farm operators or based on documented faults in the form of service reports and fault logbooks. The normal labels have

been determined by a combination of feedback of the wind farm operators, manual inspection of the data and expert knowledge.

For wind farm A all anomaly event starts were defined based on the available EDP fault logbook which only defines start timestamps for each fault. Since no further information is available, analysis of the data before every fault was used to determine possible event starts. The 'true' anomaly event starts for wind farm A can differ from the set ones.

For the wind farms B and C all starts of the anomaly events were defined based on data analysis, feedback of the wind farm operator, service report documents and expert knowledge. While the true starts of the anomaly events could potentially differ from the set ones in some cases, it is highly unlikely that the defined events start too early. If anything, anomaly event start could be earlier than defined.

The second level of labeling assigns a label to each timestamp of every dataset. These labels are called status-IDs. For the wind farms B and C they are derived from the original operating modes that were provided by the wind farm operators in combination with service report information. For wind farm A this information was not provided. In this case the status-IDs were based on the fault information from the logbook provided by EDP. For each turbine fault the preceding 14 days were marked with the status-ID 4 (fault) and the 3 days after the fault timestamp were marked with the status-ID 3 (service mode). The time ranges around the turbine fault were set with the aim in mind to reduce the risk of including anomalous behavior in the training data. As no information is available on the duration of anomalies before and after the given faults, the time ranges were chosen conservatively.

The status labels can be used to infer whether a given data point represents normal WT behavior or not. The status-IDs, their description and whether we consider the status normal, are found in table 3.3.

Status-ID	Description	Considered Normal
0	Normal operation without limitations	True
1	Derated power generation with a power restriction	False
2	Asset is idling and waits to operate again	True
3	Asset is in service mode / service team is at the site	False
4	Asset is down due a fault or other reasons	False
5	Other operational states for example system test, setup, ice build-up or emergeny power	False

3.4. Anonymization

Due to confidentiality reasons the data of wind farm B and C was anonymized. The anoymization includes the removal of all information that can directly identify the wind farms, such as the name of the wind farm, the original names of each WT, the turbine type and the location. The wind farm names were replaced by the generic names "Wind Farm A", "Wind Farm B" and "Wind Farm C" while the WT names were replaced by randomized asset IDs. However, the asset-IDs were assigned in a way that makes it still possible to link different datasets that belong to the same WT.

Additionally, the timestamps of each dataset were shifted by a random number of years. This preserves the consistency of the seasonal information, although it does distort the temporal order of the datasets.

Names of the original SCADA-features were replaced by a numeration of the features. Only features that describe power, reactive power or wind speed are recognizable by their name. Additionally, power and reactive power features have been scaled with the rated power of the turbine. This way, it is still possible to clean and analyse the data using the power curve of the WT.

All status information were aggregated from the original status data of the WTs and the name of each status condition was replaced by a number in combination with a brief description. Wind farms B and C contain detailed status information while wind farm A only contains status information which indicate turbine faults.

4. Anomaly detection evaluation

Evaluation of AD algorithms poses a difficult task. On one hand the perfect AD should detect all anomalies as soon as possible, without any false alarms, on the other hand labeling of anomalies and finding proper start and end times of anomaly events cannot be done perfectly.

Although the ground truth of every AD evaluation is almost certainly flawed [29], standard classification metrics, like the F-Score, accuracy and precision, are often used to measure the performance of AD algorithms to compare them with other algorithms or to show their overall performance [33]. The F-score in particular is widely used, but it cannot be applied to evaluate the performance of AD algorithms on normal data since true negatives are not considered in the F-score. This is one of the reasons why metrics like the F-score are not suitable for a complete evaluation.

To tackle these problems, we introduce the CARE-score in 4.1 to evaluate models on the dataset described in 3. The score is composed out of four sub-scores, each evaluating a key aspect of a good AD model. In addition to that, we conduct a mini-benchmark 4.2 to showcase the CARE-score and dataset.

4.1. The CARE-Score

In the context of AD for predictive maintenance the performance of models is often difficult to assess. To address this, we introduce the CARE-score for evaluating AD in an operational predictive maintenance setting. The CARE-score focuses on four key aspects that a good AD model for predictive maintenance should excel in, which are:

1. Coverage: Detection of as many correct anomalies as possible,
2. Accuracy: Recognition of normal behavior,
3. Reliability: Few false alarm events,
4. Earliness: Detection of anomalies before fault gets critical.