Dear Referees,

the authors would like to express their gratitude for the time and effort spend in reviewing our paper. In the attached .pdf document, the response to the referees comments # 1 and # 2 will be given.

Yours sincerely,

On behalve of all the authors,

Marcel Schedat

Enclosures:

- Response to referee comments # 1 (page 2 to 14)
- Response to referee comments # 2 (page 15 to 19)
- detailed overview of the changes made to the paper (page 20 to 55)

Response to referee comments # 1

Dear anonymous referee # 1,
Thank you very much for your feedback to improve our manuscript!
In this document, the authors' responses are added in *cursive*.

As indicated in the contribution, the work presented in this paper originates from the Master's Thesis of the first author. Based on this contribution, this seems to be very good work for a Master's Thesis. In my opinion however, the quality should be improved for this paper. More in particular, especially the presentation quality. Nevertheless, the work presented is interesting and valuable.
Throughout the comments that follow, I indicated the most important (and thus necessary to tackle) ones by (*).

General comments
(*) In general, my main concern is the lack of shown results. Without more information, it is hard to review the results and conclusions drawn. I'll give specific examples later on. A general rule of thumb could be: whenever a conclusion is drawn based on own results, it can only be checked if the results themselves are shown. So all main conclusions drawn in the paper should be preceded by the results shown in a figure or table.

Some other general comments:
- The title states that "data mining techniques" (plural) are used to model fatigue loads. However, only neural networks are used. I'd consider to use "neural networks" instead of "data mining techniques". If you want to express the comparison of multiple techniques, an adjustment of the title might be considered towards feature selection (since this is not reflected in the title at this point). In my opinion, this is not strictly necessary.

    The title was changed to "Feature selection techniques for modeling tower fatigue loads of a wind turbine with neural networks"

- Although mentioned in the abstract, the sensitivity analysis regarding the length of the data set (and the motivation for it) is not mentioned in the introduction.

    Sensitivity analysis was changed to a particular reduced set of continues data.

Specific comments
- P.1 line 19-20: it's not clear to me what you mean by "more conservative models regarding the number of features"

    With the neighborhood component analysis the number of features was reduced while maintaining the interpretability of the features used. This would not be the case of e.g. Principal Component Analysis.

    The sentence was changed.

- P.1 abstract: consider including quantitative results (e.g. errors between measured and estimated DEL do not exceed x%)

    Quite a lot results were produced for different features and operational modes. We have added absolute mean squared error for the artificial neural network model using neighborhood component analysis and the full load operational mode.

- (*) P.4 line 9-10: Some more information on the turbine would be appreciated, in particular rated power.

    This paper seeks to model tower fatigue loads of a commercial wind turbine with a rated power of 2.05 MW, a hub height of 100 m and a rotor diameter of 92.5 m in the northern part of Germany. The turbine is used by the Wind Energy Technology Institute at the Flensburg University of Applied Sciences for research purposes.

    The text has been changed and added.

- (*) P.5 line 1: More information about the measurement setup is needed. E.g. Are the measurements corrected for wall temperature before calculating bending moments? Is it possible to show the resulting (normalized) bending moments? E.g. vs time (for a smaller period) and vs windspeed

For this study, the readings from the SCADA and a load measurement system in the previously mentioned turbine were recorded over around 11 months and collected in 10 min files. The tower bottom bending are measured by strain gauges. These were installed and wired as full bridge (Wheatstone) with temperature compensation. A Wheatstone bridge is widely used in strain gauge applications because of its ability to measure small deviations in resistance. The calibration factors were determined from the results of the shunt-resistor-calibration, tower geometry and the thickness of the tower wall at the strain gauge positions (provided by the turbine manufacturer). The Offsets are determined by means of a yaw round.

We have inserted a graph showing the DELs of the tower bending moments vs time (Figure 2).

The text has been changed and added.

- P.5 line 9-11: the number used as n_eq is usually given too

The short-term damage equivalent loads for every 10 min time series were calculated. The reference number of cycles within the lifetime of 20 years were assumed to be 10^7 *cycles. The short damage equivalent load by 600 s equates to* $n_{eq} = 9.5064$. *Alternatively, the number of load cycles corresponding to 1 Hz (1Hz DEL) could be possible, but this was not decisive for the focus of the paper.*

The text has been added.

- P.5 line 14-16: How are the outliers detected? What were the (normalized) limit values to detect them?

The process of outlier detection in this study was not automated but done through visual inspection of the descriptive statistics calculated from the time series for each operational mode.

The text has been added.

- P.5 line 19: Can you give some more explanation on the descriptive statistic "mode"?

"Mode" is one of the seven descriptive statistics that we used. Mode is the most frequent value in a 10-min time series.

- P.6 Figure 1: nice figure, does help to understand the methodology

Thank you.

- p.6 line 4: This seems to be a lot of missing SCADA data. Are all of the variables missing or is it mostly due to one variable?

As described earlier in the text, measurement errors were removed. In our case this happened due to technical failure in the extraction of the SCADA and our measuring computer over approximately 2.5 months.

- P.6 line 5: This sentence is a bit confusing. The filtering by operational modes is done for the feature selection and the sensitivity analysis, isn't it?

Apologies, this is misleading. The filtering was done in general for this analysis, for the sensitivity analysis we used the operational mode of partial load since more data was available there.

- (*) P.6 line 6-8: should the mean value of ACpow be below or equal to 5kW for the datapoint to correspond to standstill? Or is it the minimum, maximum or another descriptive statistic?

In this sense, standstill corresponds to 10 min mean "ACpow" readings below or equal to 5 kW (0.25 % of nominal power); partial load to readings higher than 5 kW and below or equal to 2000 kW (97.56 % of nominal power); and full load to readings above 2000 kW.

Further explanation has been added.

- P.6 line 7-8: It is easier to interpret the limits as x% of rated power instead of x kW. Consider to change to relative values.

Percentage of nominal power was added (see "- (*) P.6 line 6-8")

- P.7 line 7-8: The sentence "Correlation coefficients above 0.95 …" deserves more explanation. Does this mean that if an explanatory feature correlates with more 0.95 to the bending moment it is considered as redundant? Or is this only the case for explanatory features among each other?

Correlation above 0.5 between the features and damage equivalent loads was considered. Correlation among all features of a particular sensor above 0.95 was considered as a redundant sensor and therefore eliminated for further analysis.

The text was changed.

- P.7, line 12: more explanation about p-value and F-statistic would be appreciated.

P-value is used as a probability measure to identify if a particular feature is significant for the outcome of the model. If a p-value is larger than 0.05 the null hypothesis is true and the feature is selected for further modelling.

The text was added.

- P.7 line 22-23 "the output is then predicted by applying a function": What kind of function?

The NCA is based on the k-NN algorithm, therefore the prediction is performed by the trained k-NN regression model.

The text has been added.

- P.7 line 27: how are the weights decided for this paper?

The weights are usually (also in the paper) assigned randomly and then adjusted by solving a minimization problem (minimizing the prediction error).

Further text has been added.

- P.7 line 28: observations = features ? This is a bit confusing, since "observations" might also be used for different measurements (in time)

The word was changed to "features".

- (*) Section 3.1: a visualization or overview of the selected and disregarded features for each dataset and technique is missing. As a reader, it is impossible to know which features were selected by which technique for which dataset except for the (few) mentioned in the text. Only by showing these results, the drawn conclusions can be checked. Moreover, it's easier to understand the conclusions if you can see the results yourself.

We have prepared a detailed overview of the three topics, which strengthens the comprehensibility:

1. *Correlation Analysis by Operation Mode*
2. *Stepwise Regression results from all operational modes*
3. *Summary of NCA for different operation modes*

We think the results are quite interesting but too much too add in the paper. We will therefore write these representations in the appendix A 1 to 3.

Appendix A 1 Correlation Analysis by operational modes
Chosen values are highlighted in red for the correlation analysis.

Feature	***Still Stand***	***Partial Load***	***Full Load***	***All Modes***
acc_x_max	*0.92*	*0.92*	*0.88*	*0.96*
acc_x_mean	*0.91*	*0.93*	*0.96*	*0.96*
acc_x_range	*0.92*	*0.92*	*0.88*	*0.96*
acc_x_mode	*0.53*	*0.08*	*0.25*	*0.19*
acc_x_std	*0.93*	*0.94*	*0.97*	*0.97*
acc_x_var	*0.77*	*0.85*	*0.94*	*0.86*
acc_y_min	*0.04*	*0.04*	*0.03*	*-0.08*
acc_y_max	*0.81*	*0.90*	*0.82*	*0.86*
acc_y_mean	*0.80*	*0.88*	*0.81*	*0.85*
acc_y_range	*0.81*	*0.90*	*0.82*	*0.86*
acc_y_mode	*0.35*	*0.14*	*0.00*	*0.06*
acc_y_std	*0.80*	*0.90*	*0.83*	*0.85*
acc_y_var	*0.70*	*0.77*	*0.82*	*0.75*
v_wind_min	*0.65*	*0.53*	*0.59*	*0.64*
v_wind_max	*0.76*	*0.90*	*0.87*	*0.80*
v_wind_mean	*0.74*	*0.83*	*0.84*	*0.74*
v_wind_range	*0.75*	*0.92*	*0.77*	*0.79*
v_wind_mode	*0.72*	*0.80*	*0.78*	*0.71*
v_wind_std	*0.73*	*0.94*	*0.81*	*0.77*
v_wind_var	*0.70*	*0.90*	*0.77*	*0.71*
v_dir_min	*0.12*	*0.08*	*0.12*	*0.16*
v_dir_max	*-0.13*	*-0.05*	*0.12*	*-0.16*
v_dir_mean	*-0.02*	*0.05*	*0.03*	*0.00*
v_dir_range	*-0.15*	*-0.07*	*0.01*	*-0.19*
v_dir_mode	*0.02*	*0.00*	*0.00*	*-0.01*
v_dir_std	*-0.15*	*0.07*	*0.11*	*-0.11*
v_dir_var	*-0.11*	*0.04*	*0.10*	*-0.08*
omega_gen_min	*-0.08*	*0.55*	*-0.79*	*0.64*
omega_gen_max	*0.19*	*0.80*	*0.81*	*0.68*
omega_gen_mean	*0.04*	*0.71*	*0.08*	*0.67*
omega_gen_range	*0.38*	*0.37*	*0.88*	*0.28*
omega_gen_mode	*0.02*	*0.65*	*-0.05*	*0.66*
omega_gen_std	*0.33*	*0.29*	*0.94*	*0.19*
omega_gen_var	*0.30*	*0.22*	*0.93*	*0.13*

Feature	*Still Stand*	*Partial Load*	*Full Load*	*All Modes*
air_density_min	*-0.10*	*0.22*	*0.05*	*0.23*
air_density_max	*-0.11*	*0.22*	*0.06*	*0.23*
air_density_mean	*-0.10*	*0.22*	*0.05*	*0.23*
air_density_range	*-0.02*	*0.01*	*0.08*	*-0.07*
air_density_mode	*-0.10*	*0.22*	*0.05*	*0.23*
air_density_std	*-0.02*	*0.00*	*0.09*	*-0.07*
air_density_var	*-0.02*	*0.02*	*0.07*	*-0.03*
pitch_min	*0.27*	*0.03*	*0.71*	*-0.35*
pitch_max	*0.41*	*0.31*	*0.87*	*-0.25*
pitch_mean	*0.35*	*0.21*	*0.81*	*-0.31*
pitch_range	*0.30*	*0.31*	*0.55*	*0.37*
pitch_mode	*0.36*	*0.12*	*0.66*	*-0.32*
pitch_std	*0.30*	*0.24*	*0.32*	*0.25*
pitch_var	*0.26*	*0.15*	*0.28*	*0.09*
ACpow_min	*-0.15*	*0.64*	*-0.05*	*0.82*
ACpow_max	*0.20*	*0.89*	*0.83*	*0.89*
ACpow_mean	*0.04*	*0.81*	*0.29*	*0.88*
ACpow_range	*0.21*	*0.92*	*0.17*	*0.70*
ACpow_mode	*0.00*	*0.75*	*-0.15*	*0.85*
ACpow_std	*0.21*	*0.88*	*-0.07*	*0.59*
ACpow_var	*0.20*	*0.70*	*-0.12*	*0.45*

Appendix A 2 Stepwise Regression results for different operation modes
Chosen values are marked with an "x".

Feature	Still Stand	Partial Load	Full Load	All Modes
acc_x_min		x		x
acc_x_max	x	x	x	x
acc_x_mean	x	x	x	x
acc_x_range				
acc_x_mode	x	x	x	x
acc_x_std	x	x	x	x
acc_x_var	x	x	x	x
acc_y_min				
acc_y_max	x			x
acc_y_mean	x	x	x	x
acc_y_range	x			
acc_y_mode			x	
acc_y_std		x		
acc_y_var	x	x		x
v_wind_min		x	x	x
v_wind_max				
v_wind_mean	x	x	x	x
v_wind_range		x		
v_wind_mode	x		x	x
v_wind_std	x	x		x
v_wind_var	x	x	x	
v_dir_min	x			
v_dir_max	x	x		
v_dir_mean	x	x		x
v_dir_range				x
v_dir_mode				
v_dir_std	x	x	x	x
v_dir_var	x	x		x
omega_gen_min	x	x	x	
omega_gen_max		x		
omega_gen_mean	x	x		
omega_gen_range	x			
omega_gen_mode	x	x		x
omega_gen_std	x		x	x
omega_gen_var	x	x	x	x

Feature	*Still Stand*	*Partial Load*	*Full Load*	*All Modes*
air_density_min			*x*	
air_density_max				
air_density_mean				
air_density_range	*x*			
air_density_mode			*x*	
air_density_std	*x*	*x*		*x*
air_density_var	*x*	*x*		
pitch_min	*x*	*x*	*x*	*x*
pitch_max		*x*		*x*
pitch_mean		*x*	*x*	*x*
pitch_range				
pitch_mode	*x*	*x*	*x*	*x*
pitch_std		*x*	*x*	*x*
pitch_var	*x*	*x*		*x*
ACpow_min		*x*		*x*
ACpow_max	*x*	*x*	*x*	*x*
ACpow_mean	*x*	*x*	*x*	*x*
ACpow_range				
ACpow_mode	*x*	*x*		*x*
ACpow_std	*x*	*x*		*x*
ACpow_var	*x*	*x*		*x*

Appendix A 3 Summary of NCA for different operation modes
Chosen values are marked with a "x".

Feature	***Still Stand***	***Partial Load***	***Full Load***	***All Modes***
omega_min				
omega_max				
omega_mean				
omega_range				
omega_mode				
omega_std				
omega_var				
acc_x_min				
acc_x_max	x			
acc_x_mean	x	x	x	x
acc_x_range	x			
acc_x_mode				
acc_x_std	x	x	x	x
acc_x_var				
acc_y_min				
acc_y_max	x			x
acc_y_mean				
acc_y_range	x			x
acc_y_mode				
acc_y_std	x			
acc_y_var				
v_wind_min				
v_wind_max	x			
v_wind_mean		x		x
v_wind_range				
v_wind_mode				
v_wind_std		x		x
v_wind_var				x
v_dir_min				
v_dir_max				
v_dir_mean	x			x
v_dir_range				x
v_dir_mode	x			
v_dir_std	x	x		
v_dir_var				
omega_gen_min	x			
omega_gen_max	x	x		

Feature	*Still Stand*	*Partial Load*	*Full Load*	*All Modes*
omega_gen_mean				
omega_gen_range	x	x		x
omega_gen_mode				
omega_gen_std			x	
omega_gen_var				
air_density_min			x	
air_density_max				
air_density_mean				
air_density_range				
air_density_mode				
air_density_std				
air_density_var				
pitch_min			x	
pitch_max				
pitch_mean			x	
pitch_range		x		
pitch_mode				
pitch_std			x	
pitch_var				
ACpow_min		x		
ACpow_max				
ACpow_mean		x		x
ACpow_range	x	x		x
ACpow_mode				
ACpow_std	x		x	x
ACpow_var				

- P.9 line 19: a figure showing results about the collinearity would be helpful for this discussion

All results of the correlating analysis were attached in the appendix A 1.

- P.9 line 22-23 “Many of the remaining variables … model”: an example with specific resulting values would be helpful

Same answer as for “P.9 line 19”

- P.10 line 6: How does stepwise regression avoid multicollinearity?

Stepwise regression does not avoid collinearity directly. In Stepwise regression features are added to the regression model one by one. If a feature added results in a better model (meaning that the prediction accuracy is better) then the feature is kept. Generally, a feature which is collinear with others would not improve the accuracy of a model and there it is dropped by the stepwise regression.

In order to avoid misunderstandings, we have removed this sentence.

- P.11 Section 3.1.2: Discussion about PCA seems to be missing

For each operational mode PCA was performed to account for potential collinearity in the feature-set. This was done consistently with an explained variance of 99 % remaining.

The text has been added.

- (*) P.12 Section 3.2.1: I like the idea to first give more precise results for one neural network model before comparing all of them. However, more information and results should be given. Which were the final features used for this one model? How did the training, testing and validation data look like (for example plot of the measured normalized power curve for all three datasets, can be based on mean statistics)? Plots of the measured and predicted DEL vs mean windspeed, errors vs mean windspeed for example.

We have rebuilt the entire paragraph to give a better understanding on how precise the results (predicted DELs) in comparison to the measured DELs are. For example, in Figure 5 you can see predicted and the calculated (measured) DELs over a number of 10 min time series and the corresponding prediction error. Furthermore, Figure 6 gives the reader an impression of the behaviour prediction error vs wind speed.

- P.12 Section 3.2.1: some of the information and results given seem irrelevant for this discussion. E.g. at which epoch the training stopped. Consider to omit this from the paper or to include it in the discussion to show the added value

We agree with your comment.

We removed this part and rebuilt the entire paragraph.

- P.12 Figure 3: Personally, I don't think this plot adds value to the paper. If not needed for main conclusions, consider to remove it.

We agree and have therefore removed it.

- P.12 line 11-13: these datapoints (outliers) are not visible in the figure. Maybe consider a different visualization

We removed the figure.

- P.12 line 10-11 and p.13 Figure 4: Personally I don't see the added value of this figure and discussion.

We removed this part and rebuilt the entire paragraph.

- (*) P.12 line 12 and further: didn't you exclude outliers from your dataset? Did you take a closer look to the time signals of bending moment and different SCADA parameters to check what is causing these high errors?

The mean error in percent is calculated as follow:

$$mean\ error = \frac{(calculated\ DEL - predicted\ DEL)}{calculated\ DEL} * 100$$

High prediction errors are expected when the DEL is low. For example, the calculated DEL is 30 kNm and the predicted is 150 Nm then we will have an error of -400 %. In terms of prediction error it is extremely high, but we are only 120 kNm off. If we are 120 kNm off in full load then this error can be less then 1 %.

We added an additional plot as suggested to see the high error occurring to low wind speed (Figure 6).

- P.12 line 16: I cannot find the result 0.99486 in the figures or tables

All models were run again based on better comparability with other studies and the comments made above.

- P.14 Figure 5: It's not clear to which data sets the figures exactly correspond. Add sublabels please.

A*ppendix A 1 to 3 was added for correlation, stepwise regression and NCA for different operational modes.*

Additionally the whole chapter was rewritten.

- Section 3.2.1: An additional conclusion could be made: From the introduction I understood Vera-Tudela and Kühn did a similar analysis with slightly different techniques. How do your results compare to theirs?

Results were changed to "mean error in %" and "standard deviation of the error in %" to make it comparable to the results from Vera-Tudela. The complete analysis was redone therefor changes in numbers are possible.

A short comparison was added in Section 3.2.2. in the paper:

"The results from the full load model can be compared to existing work from Vera-Tudela and Kühn (2014) where the mean error is below 0.22 for all feature-sets. The maximum absolute error and standard deviation of the error also confirm the results. The accuracy of the results from the partial load model is slightly worse for all features-sets."

Additional sentence added in the conclusions: "It can be concluded that the performance of NN is influenced by the operational mode that the WT. The highest accuracy was achieved when the WT was operating in full load and the lowest in stand still."

- P.14 Section 3.2.2: Additional figures might be helpful here too. For example, measured and predicted (by the different models) DEL vs mean windspeed, where a different color is used for each model/operational mode.

We added all relevant information on Figure 5 and 6. For all operational modes the calculated DELs vs. predicted DELs highlights the accuracy of the model and the improvement as the WT operates.

- P.14 Figure 6: if you want to show DELs are lower for standstill than during full load, it is much easier to plot them on the same graph. Moreover, are the results shown here all test data? Why don't you show all test data instead of only 100 points for each operational mode?

This was intended to highlight that values close to 0 can have a high error in %. The figure 6 was removed and replaced by an overview of the measured DEL vs predicted DEL (figure 5).

- P.16, Section 3.3: What is the exact intention of this analysis? Is it to determine the minimum period needed to measure the bending moments? If that's the case, shouldn't the focus be on the dataset containing the least data? To make sure a good estimation is obtained for that operational mode too?

We changed the section by using the first 50 % of the data available in partial load for training the ANN and tested it on to the remaining 50% of the data.

- P.16 line 11: is the dataset increased consecutive in time or by randomly picking data from the entire dataset of one year?

No, for the previous sensitivity analysis we also used always the first x percent of data consecutive.

- Section 3.3: An additional conclusion could be made: From the introduction I understood Smolka and Cheng did a similar analysi. How do your results compare to theirs?

In the study of Smolka and Cheng only the correlation analysis for selecting features was conducted. In our paper we compared different methods and we were able to reduce the features significantly with NCA. This paper conducted an additional analysis for the standstill mode where significant increase of mean error was identified.

- (*) p.17 line 11-12: The first part of this conclusion is not clear to me. What do you mean with "conservative model regarding the number of features"? The second part doesn't seem to be true. Looking at Tables 2 and 3, the lowest mean errors are rarely found for NCA.

The results indicate that using all data and applying neighborhood component analysis for feature selection yields an interpretable and low dimensional feature-set while maintaining high accuracy.

The text has been changed.

- P.18 line 6-10: If the purpose is to eliminate the need for installing strain gauges on every turbine, it seems it is especially necessary the model is validated on a different turbine. Training the model with data from multiple turbines might not be necessary.

To be able to generalize the results obtained from this study, the NN model requires validation with data collected from a different wind turbines with the same specifications.
The text has been changed.

- (*) P.18 line 12: Why would you want to train the neural network with larger datasets? Wasn't one of your conclusions that you didn't need as much data to have a model equally accurate?

This is referred to the other operational mode such as stand still and full load. We had limited data there that is why we used partial load to test the model on 50% of the data.
The text has been changed.

Technical corrections

The comments hereafter are not critical and are meant to improve the readability of the paper.

- p.1 line 12-15: very long sentence, consider to split it up

done

- p.3 line 1-9: different lengths of datasets were used for the different analyses shown in this paper. The summary written here seems to suggest 2,5 months of data was used to model the thrust loads. However, only 2 weeks were used to train the model and (in case of operational data) one year was used to validate. On the other hand, 2,5 months of data was used to perform a Pearson correlation analysis.

corrected

- P.3 line 26-32: If I understood it correctly, your work is actually similar to the work of Seifert et al, except you have done it for tower bending moments, while Seifert et al. did it for blade root bending moments.

Our focus was to identify the required feature for build such a model and how or of these effects the accuracy. This was performed by several feature selection techniques. We could conclude, that NCA was able to identify the most relevant features without linear transformation by i.e. PCA while maintaining a high accuracy of the model.

- P.6 line 4: in my opinion, 6044 hours is not easier to interpret than 36266 observations of 10 minutes. I think "a little over 8 months" would be better. Similarly in the conclusion (p.18 line 3-5)

done

- P.6 line 4: Considered to add also the percentage of remaining data after the removal of missing data.

done

- P.8 line 7: typo "squared"

done

- P.9 line 25: It might be helpful to clearly state that from this point the second technique, PCA, is discussed.

done

- P.10, Figure 2: add a line at 99% to increase visibility
  You can see the relevant point in Figure 3 (Figure 2 was changed).

- P.11 line 8: typo "datasets is the again the"
  c*orrected*

- P.13 line 5: sentence can be missed very easily
  The paragraph was completely rewritten.

- P.14 line 10: reference to results is missing
  has been added

- P.16 line 14: an equivalent number in time (weeks, months) is easier to interpret than 10241 observations
  The paragraph was completely rewritten and the comments were noted.

Response to referee comments #2

Dear anonymous referee # 2,
We thank you for providing valuable comments to improve our manuscript.
In this document, the authors' responses are added in *cursive*.

Received and published: 10 February 2020

The paper deals with the estimation of damage equivalent tower bending moments from SCADA data using a neural network approach. In particular it focuses on methods for feature selection to determine which of the available parameters has to be included as input to the network. The work is interesting and worth publication in general. However, improvements are necessary. The paper should also be revised carefully to improve the presentation quality.

General comments
The title implies that several data mining techniques for modeling tower fatigue loads are compared. However, it seems that the main work focuses on feature reduction techniques. In my opinion the title should be adjusted to better reflect the content of the paper.

A comprehensive literature survey has been undertaken to investigate state of research related to the estimation of loads from SCADA data. However, the paper also focuses on methods for feature reduction. Are there studies outside of wind energy which compare feature reduction methods? If so, how do the results compare to those presented in this paper?

Specific comments
p.1, l.18-20: What is meant by "conservative" models?

This expression was disadvantageous and was therefore changed to "The Neighborhood component analysis yields the minimum number of features required while maintaining the interpretability with an absolute mean squared error of around 2 % for full load."
In this way, the word "conservative" meant a low dimensional feature-set.

p.1, l.22: What do you mean with "deployment"? It may not be the correct word. How does that relate to the competition?

The expression has been replaced by "due to the rapid growth of the wind energy installed capacity".

p.2,l.7: Could you please explain why load measurement systems are required for this purpose? The structural condition is usually monitored with accelerometers.

For us, the focus was on determining the bending moments on a tubular steel tower of a wind turbine. A continuous measurements and recording of the strain (in this case by strain gauges) were important for this. The acceleration sensors on the tower used in commercial systems are primarily used for vibration monitoring (mostly in the area of the tower head). If a defined limit value is reached, a warning or alarm follows with a corresponding reaction of the control unit (e.g. shutdown). The continuous decoupling of signals from these acceleration sensors was not readily available to us at the time of the measurement campaign.

p.4,Table 1: From the rest of the paper it seems that the dependent variables for the modeling are the fore-aft bending moment and not strains (see p.5, l.11). Please clarify.

The tower bending moments are calculated by the measured strains with the help of strain gauges (Chapter 2.1 refers to this calculation of the moments)

p.5, l.2: Please explain what you mean by "short-term" equivalent load?

The short-term DEL is based on a 10 min time series without extrapolation to its lifetime. In our chase we used for 10 min $n_{eq} = 9.5064$*, which is equivalent to* 10^7 *cycles in 20 years. Alternatively, the number of load cycles corresponding to 1 Hz (1Hz DEL) could be possible, but this was not decisive for the focus of the paper.*

The information has been added to the text.

p.5, l.2: Please explain what you mean by "short-term" equivalent load?
Repetition of the question before.

p.5, l.11: Why m=3?
We determined the DELs on a tubular steel tower. For clarity, we have chosen only one inverse slope m = 3 for steel (m = 4 or 5 are common, too).

p.7: It is nice that the four approaches are explained briefly. However, it should also be discussed how the approaches compare with respect to feature reduction. What are their limitations or advantages? Maybe the methods can also be compared using a table which may be easier to follow for the reader?
The results of the methods are now attached to the appendix A 1 to 3. The reader can see which features were selected for correlation analysis, stepwise regression and NCA.

p.8, l.8: How does the validation subset generalize the transfer function? And what is meant be "transfer function" anyhow? Is that the network itself?
The transfer function is part of the ANN. The weights are adjusted and later a weighted sum is passed through the activation function. The validation set is used to "test" the prediction. If the prediction has a bad accuracy but the training data performs well it is overfitted and the weights are readjusted. The expression "transfer function" is changed to "prediction model".

p.8, l.13: The choice for the NN architecture should be based on some rational and not on a default suggestion by a software. Why do you think that the chosen NN architecture is reasonable? Can you give another justification than that this is the default setting of a software?
There is no rule on how to choose the hyperparameters for a black box model like ANN. They can be chosen by default or randomly and later adjusted based on the performance of the ANN. In this case there was no need to adjust the parameters as the model performed well. We tried varying the parameters (which is not part of the paper and requires a separate analysis) and got similar results. It could be possible to investigate on the number of hidden layers required to achieve certain results or the learning rates which would be a separate analysis.

p.8, l.15: See comment above.
Please see answer above. Our focus was not on optimizing the architecture, which is why we kept the default setting.

p.9, l.14: The DELs are calculated for the bending moments and not for the thrust.
Right, t*he DELs are calculated for the fore-aft moments of the wind turbine tower.*
We changed the text.

p.9, l.15: What do you mean by "facing the wind"? Are measurement with large yaw errors included in the dataset? And how can the wind not affect the thrust load?
In this investigation, we neglected a potential yaw angle error on the system and only calculated the tower bending moment in the fore-aft direction at the base in alignment with the nacelle. The investigation of the influence of inclined wind flows, e.g. due to yaw errors, were not part of this work and is currently under further investigation.
The sentence in text has been deleted.

p.9, l.23ff: Is PCA not applied? Or does the rest of the page refer to PCA?
The page refers to 11 months of data used to analyse. One of the feature selection techniques is the correlation analysis, where on top PCA is applied for dimensionality reduction.
We added this information to the text.

p.10,l.19-21: The paper is justified in Chapter 1 by the fact that NCA was not investigated in other studies so far. It is therefore a bit disappointing, that there are only two sentences related to this method in this section, especially considering that results from the other approaches are described in much greater detail.

NCA is described on p. 26 (this document) as a feature selection technique. The technique itself is no particularly novel there for a brief introduction seemed appropriate. We added a complete table on the feature selection techniques and their results. The results from NCA can be found in appendix A 3.

p.11,l.17: Isn't it more a change in wind speed that has an effect on rotor speed and tower kinematics alike? Do you mean with this sentence that rotor speed and tower defections are correlated?

The wind speed influences the rotational speed and the rotational speed is correlated with the DEL. If there is no rotation at the wind turbine we still have a changing DEL due to the wind speed and therefore the wind speed is higher correlated with the DEL then only the rotational speed.
Here you can see our correlations (appendix A 1):

- *correlation of mean wind speed (v_wind_mean) with DEL: 0.74*
- *correlation of rotational speed (omega_gen_mean) with DEL: 0.67*

They do not exclude each other but complement.

p.11,l.22-25: This is surprising. Was it investigated in detail? If there is no correlation, the stepwise regression should not select this feature. Is there an error in the stepwise regression approach or is there another explanation for the pitch angle?

If the stepwise regression does not select a particular feature, this does not mean that it is not important or correlated. If a feature does not improve the prediction accuracy if will not be selected by stepwise regression. This can happen when another feature is already selected in a model which provides similar information. In this way multicollinearity is avoided.

p.11: Was PCA not applied?

PCA was applied and this information was added to the text.

p.14, l.16ff: It is not shown in Figure 6 that DELs during standstill are lower as the DELs are normalized. Is there another way to illustrate it?

This chapter has been completely revised in large parts and the new illustrations provide an improved overview.

p.16, l.1: What do you mean by good practice? It seems that feature-selection has no impact on the results. This is contradicting to the research by Sharma and Saroha (p.6) which states that feature-selection should result in more accurate results. Could you please discuss?

Features selection method helped to reduce the amount of information needed to build a particular model. Feature selection reduces the needed information building a particular model since some of the features do not provide any sufficient information to model the DEL. This ultimately leads to a reduced amount of sensors required modelling the DEL (Appendix A1 to 3 added for the feature selection results on correlation analysis, stepwise regression and NCA).

p.16, table 3: It seems that feature selection has not impact on the estimation results from the NN. Can this indicate that there are still too many features? Usually, each method can be varied by changing some parameters. Have you conducted a parameter study to investigate, if more strict settings for feature selection would result in even smaller feature sets without loosing accuracy?

This indicates that the feature selection techniques chose the features which has sufficient information to later construct such a model. Investigating the parameters for each feature selection technique is unfortunately not in the scope of this study. It could be possible to reduce the features even further and maintaining the same accuracy, this would need further investigation.

p.16, l.13f: Please discuss how this result can be used to reduce time and costs for data collection in practice. To my understanding the 40% of the data was randomly selected out of 1 year of measurements. That means that is still requires to measure for one year. To me, one year of measurements sounds reasonable to cover all operational conditions, seasonal variations, etc. I cannot see how that can be reduced really. If still 1 year of measurement is required the purpose of this study remains unclear. Please give a justification why this study was undertaken and why results should be presented in this paper.

We changed the sensitivity study to a particular data set (50% of partial load). This were not chosen randomly, they refer to the first 50% of the data. We use the model to predict "future" DELs and investigate how well the model performs. This would mean that collecting around 3 month of data to calculate the DEL would be sufficient to predict it for the next 3 month (for partial load).

p.17,l.11f: I cannot see from table 3 that a NN based on NCA is superior in terms of accuracy. In the abstract it is also mentioned that all NN result in similar accuracy. See also p.15, l. 5ff

It is superior in term of least number of features required while keeping the interpretability of the features. This section was completely rewritten.

Technical comments

p.1, l.17: "with the partial load model"

corrected

p.1, l.25: "failures for example"

corrected

p.1, l.29: Are there two spaces "of WTs"?

corrected

p.1, l.29: "can potentially be"

corrected

p.2, l.8: I don't think that "sophisticated" is the correct word. Maybe "challenging"?

corrected

p.2, l.10: I don't think that "briefly visited" is the correct wording.

corrected

p.2,l.17-34: This section contains quite general statements and also the motivation for performing fatigue load estimation. Why is it placed in the middle of the literature review? Should it be moved to the beginning of chapter 1?

corrected

p.3, l.34: I don't think that "scarce" is the correct wording.

corrected

p.5, l.13: It is not the method for "development of the paper" which is shown.

corrected

p.6, l.11: "most" instead of "more"?

corrected

p.6, l.11: "For this purpose"?

corrected to: "For the learning process"

p.7, l.23: "differentiate" instead of "differ"?
corrected

p.8, l.23: "sought"? What does that mean?
The Chapter has been revised. The word

p.8, l.23: "accurately" instead of "appropriately"?
corrected

p.8, l.21-23: The same is written just a few lines above.
corrected

p.9, l.3-5: This sentence is hard to understand. Could you please formulate it in a different way?
Corrected to: "The results indicate that the accelerations in both directions (i.e. x and y-axis) are highly correlated with the DELs. The standard deviation of the acceleration in the x-direction presents the highest correlation with a coefficient of 0.97, depicting an almost linear relationship between this feature and the dependent variable."

p.12, l.15: "estimated DELs" instead of "trained DELs"?
Chapter has been revised

p.14, l.10: "It can be observed"
corrected

Feature selection techniques for ~~M~~mode~~l~~ling tower fatigue loads of a wind turbine ~~using~~ with neural networks~~data mining techniques on SCADA data~~

Artur Movsessian[1], Marcel Schedat[2], Torsten Faber[2]

[1]Institute for Infrastructure and Environment, School of Engineering, University of Edinburgh, United Kingdom
[2]Wind Energy Technology Institute, University of Applied Sciences Flensburg, Flensburg, 24943, Germany

Correspondence to: Marcel Schedat (marcel.schedat@hs-flensburg.de)

Abstract. The rapid development of the wind industry in recent decades and the establishment of this technology as a mature and cost-competitive alternative have stressed the need for sophisticated maintenance and monitoring methods. Structural health monitoring has risen as a diagnosis strategy to detect damage or failures in wind turbine structures with the help of measuring sensors. The amount of data recorded by the structural health monitoring system can potentially be used to obtain knowledge about the condition and remaining lifetime of wind turbines. Machine learning techniques provide the opportunity to extract this information, thereby improving the reliability and cost-effectiveness of the wind industry as well. This paper demonstrates the modeling of damage equivalent loads of the fore-aft bending moments of a wind turbine tower ~~with~~ highlighting the advantage of using the neighborhood component analysis. ~~as a~~This feature selection technique ~~in~~ is ~~comparison~~ compared to common dimension reduction/feature selection techniques such as correlation analysis, stepwise regression or principal component analysis. For this study, ~~a one-year measuring period~~recordings of data ~~was~~ were gathered during a period of approximately 11 months, pre-processed, and filtered by different operational modes, namely stand~~-~~still, full load, and partial load. The results indicate that all feature selection techniques were able to maintain a high accuracy when trained with artificial neural networks. The neighborhood component analysis yields the lowest number of features required while maintaining the interpretability with an absolute mean squared error of around 0.07 % for full load. Finally, the applicability of the resulting model for predicting loads in the wind turbine is tested by reducing the amount of data used for training. This analysis shows that the predictive model can be used for continuous monitoring of loads in the tower of the wind turbine.~~Finally, a sensitivity analysis was performed in the partial load model to determine the required length of the data collection campaign that guarantees the most precise results. The results indicate that applying neighborhood component analysis yields more conservative models regarding the number of features and equally accurate outcomes than traditional feature selection techniques.~~

1 Introduction

Wind power is becoming the electricity-generati~~on~~ng technology with the lowest costs in several areas of the world (REN21, 2018). Given the fierce competition in the industry due to the rapid growth of wind energy installed capacity globally~~large~~

~~deployment of wind turbines (WTs),~~ it is important to continuously look for alternatives to make this technology more cost-effective. The possibility of monitoring with sensors ~~or sensor systems~~ has enabled the gathering and ~~supervising~~ supervision of data regarding the ~~about an object's~~ condition of a structure to, e.g. ~~-~~detect ~~for example~~ failures. Particularly, structural health monitoring (SHM) ~~at~~ in wind turbines (WTs) allows ~~to monitor~~monitoring the structural behaviour and stresses of structures such as blades, towers, and foundations.

While machine learning techniques are widely applied in industries such as the automotive, information technology and communication, the wind industry is starting to explore the suitability of these promising methods for its benefits. Although data-driven attempts have been made to estimate the loads acting on the turbine using available information from the SCADA system, there is no consensus yet on the type of a relationship existent between these data and actual load measurements. In the last years, the focus on this topic increased. This section aims to review available scientific literature regarding modeling loads with existing SCADA data for WT.

SHM systems could be used to verify structural safety and determine the remaining useful lifetime (RUL) of WTs (Schedat et al., 2016). Moreover, information gathered through SHM during the lifetime of WTs can ~~be~~ potentially be used to identify structural weaknesses and feed this information back to the manufacturers, ultimately improving the design of new turbines (Ziegler et al., 2018). Another potential benefit of SHM is a decrease in maintenance costs. Typically, operation and maintenance costs (including both fixed and variable costs) represent approximately 20 to ~~-~~25 % of the total levelized cost of electricity (LCOE) (IRENA, 2015). SHM could reduce this share by allowing the implementation and set in place of more efficient maintenance practices such as predictive maintenance while enabling ~~a~~ better spare-parts inventory management. Consequently, downtime is reduced, and production is increased.

Currently, the assessment and evaluation of the structural condition of WTs without a load measurement system can be challenging. Particularly, the estimation of fatigue loads can be ~~frequently sophisticated~~difficult due to a lack of information (Melsheimer et al., 2015; Schedat and Faber, 2017). Therefore, exploring the ways to mine data from SHM systems and extract valuable information becomes an interesting and ~~briefly visited~~ high demanded field of research.

The reconstruction or estimation of loads using statistics from SCADA data ~~were~~ was already presented and tested in the mid-2000s. Cosack and Kühn (Cosack and Kühn, 2006) developed a stepwise regression model for estimating the rotor thrust. Despite the good results (i.e. deviations between the calculated and the estimated loads ranged from 5.4% to ~~-~~ 7.3% in the worst case), the presented model was too complex and time-consuming with further restrictions. In a new development of the model, an estimation method for the corresponding target values (damage-equivalent loads and the load magnitude distributions) used neural networks (Cosack, 2010; Cosack and Kühn, 2007).

~~While machine learning techniques are widely applied in industries such as the automotive, information technology and communication, the wind industry is starting to explore the suitability of these promising methods for its purposes. Although data-driven attempts have been made to estimate the loads acting on the turbine using available information from the SCADA system, there is no consensus yet on the type of a relationship existent between these data and actual load measurements. In~~

~~the last years, the focus on this topic increased. This section aims to review available scientific literature regarding modeling loads with existing SCADA data.~~

Ziegler et al. (Ziegler et al., 2018) have recently performed a literature review and have assessed the development of the lifetime-extension market of onshore WTs. The alternative of extending the lifetime of a WT, as opposed to repowering or decommissioning, is appealing given the potential increase of returns on investments (ROIs), however, not much public research has been done on this matter. The authors contributed, then, by comparing updated load simulations and inspections for lifetime extension assessments in Germany, Spain, Denmark, and the United Kingdom. Particularly, for lifetime extension to be a feasible alternative, the structural integrity of the turbine should not compromise the level of safety. In this regard, the survey performed by the authors determined that, apart from the use of SCADA systems, no short-term load measurements or monitoring are carried out in the countries surveyed (a few exceptions were identified in the UK, where load reassessment is performed). They found that most interviewees focus on practical assessments for cost reasons. Nevertheless, these practical inspections are no guarantee that the safety level can be maintained during the lifetime extension. The authors concluded that new ~~O&M~~operation and maintenance strategies and data-processing methodologies are necessary for lifetime extension purposes. Moreover, data-driven approaches may contribute to cost reduction of lifetime extension assessments.

In line with the findings of Ziegler et al. (Ziegler et al., 2018), other authors have worked on the aforementioned data-driven approaches. Noppe et al. (Noppe et al., 2018), for example, reconstructed the thrust loads history of a WT based on both simulated and measured SCADA data. The data gathered corresponded to operational 1 s and 10 min ~~data for 2.5 months~~. Moreover, the data is segregated into different operational modes. The selection of explanatory variables that the authors performed was based on a Pearson correlation analysis. The first two weeks of operational data were used to model ~~T~~the thrust loads ~~were modeled~~ using neural networks and validated by one year of data. The model has the following input features: wind speed, blade pitch angle, rotor speed, and generated power. The results of this paper showed that the constructed model was able to estimate thrust loads with a relative error that does not exceed 15 %. The authors also concluded that the use of simulated data yielded slightly better results and that adjustments in the hyperparameters of the neural networks had no significant impact on the estimated thrust loads.

Relatedly, Vera-Tudela and Kühn (Vera-Tudela and Kühn, 2014) focused on the selection of variables to be used for fatigue load monitoring and attempted to define an optimum set of explanatory variables for that purpose. The authors identified 117 potential variables (13 statistics of 9 SCADA signals) used in related scientific literature. Among them, the mean of generator speed, electrical power, and pitch angle have been the most commonly used. The authors decided to apply several feature selection methods to six sets of variables. The methods chosen included Spearman coefficients, stepwise regression, cross-correlation, hierarchical clustering, and principal components. To evaluate the outcomes of the feature selection methods a feedforward neural network was employed. The authors concluded that principal components yielded the best set of variables, however, the resulting set lost expertise knowledge about the relation between the variables. In this sense, ~~a combination of~~ ranking the variables by their corresponding Spearman coefficients resulted in a fair compromise between the number of features required to monitor the damage equivalent load for blade out of plane bending moment and the available expert~~ise~~ knowledge.

Smolka and Cheng (Smolka and Cheng, 2013) examined the amount and type of data necessary to determine a fatigue estimator for the operational lifetime of a WT. The inputs for the neural network are selected by means of a correlation analysis applied to standard data statistics of available SCADA signals such as electrical power, generator speed, pitch angle, among others. The authors concluded that the minimum training data sample size required is approximately half a month worth of measurements.

Seifert et al. (Seifert et al., 2017), acknowledging the complexity and cost of handling extra measurements, assessed the minimum needed size of a training sample to predict fatigue loads using 10 min statistics of SCADA signals and neural networks. In a sense, Seifert et al.'s work is an extension or continuation of Vera-Tudela and Kühn's (Vera-Tudela and Kühn, 2014) and Smolka et al.'s (Smolka and Cheng, 2013). Seifert et al. (Seifert et al., 2017) tested different sample sizes~~, like a k-fold cross-validation,~~ varying between one day (i.e., 144 records) and four months (i.e., 4032 records) of measurements. They determined that a sample of 2016 records of 10 min statistics are sufficient to predict flap wise blade root bending moments of a WT independent of seasonal effects.

Artificial neural networks can only perform as good as the information provided to them, thus the features used to train them are key to obtain high accuracy in the results with a parsimonious model. So far, little ~~Scarce~~ research has been done regarding feature selection for mode~~l~~ling tower fatigue loads. The available literature has focused on techniques such as correlation analysis, principal component analysis (PCA) and stepwise regression to select the best subset of information. This paper aims to assess the use of Neighbourhood Component Analysis (NCA) as a feature selection technique to extract relevant information from SCADA data in order to train artificial neural networks and model fatigue loads.

The paper is organized as follows: S~~s~~ection 2 outlines the applied methodology in this study, section 3 summarizes the results, and, finally, section 4 presents the conclusions derived from the obtained results.

2 Data and Methodology

2.1 Wind turbine and SCADA data

This paper seeks to model tower fatigue loads of a commercial wind turbine with a rated power of 2.05 MW, a hub height of 100 m and a rotor diameter of 92.5 m in the northern part of Germany. The turbine is used for research purposes by the Wind Energy Technology Institute at the Flensburg University of Applied Sciences.~~This paper seeks to model tower fatigue loads of a real WT located in Schleswig-Holstein, North Germany. The turbine is used by the Wind Energy Technology Institute at the Flensburg University of Applied Sciences for research purposes.~~ For this study, the readings from the SCADA and a load measurement system in the previously mentioned turbine were recorded for around 11 months and collected in 10 min files. The tower bottom bending is measured by strain gauges. These were installed and wired as full bridge (Wheatstone) with temperature compensation. A Wheatstone bridge is widely used in strain gauge applications because of its ability to measure small deviations in resistance. The calibration factors were determined from the results of the shunt-resistor-calibration, tower geometry and the thickness of the tower wall at the strain gauge positions (provided by the turbine manufacturer). The offsets are determined by means of a yaw round. ~~For this study, the readings from the SCADA and strain gauge sensors in the previously mentioned turbine were recorded over one year and collected in 10 min files.~~ The sensors used to extract features for the model are described in Table 1 and were selected based on ~~the~~ a literature review and consultations with an application engineer.

Table 1 – Description of SCADA sensors selected

Feature name	***Description***	***Unit of measurement***	***Frequency [Hz]***
	Explanatory variables		
Omega	Rotational speed at the rotor	rpm	20
acc_x	Acceleration fore-aft (x-direction)	mm s^{-2}	20
acc_y	Acceleration side-side (y-direction)	mm s^{-2}	20
v_wind	Wind speed	m s^{-2}	20
v_dir	Relative wind direction	degree	10
omega_gen	Rotational speed at the generator	rpm	20
air_density	Air density	kg m^{-3}	20
Pitch	pitch angle	degree	20
ACpow	Active power output	kW	20
	Dependent variables		
Bieg1_060_240	Gauge sensor located at 60 & 240 degree inside the tower bottom	kNm	50
Bieg2_150_330	Gauge sensor located at 150 & 330 degree inside the tower bottom	kNm	50

The strain gauges measurements at the turbine were transformed into a resultant fore-aft tower bending moment, which was later used to calculate the short-term damage equivalent load (DEL) for every 10 min time series. This transformation was performed by means of a rainflow counting algorithm and later, the resulting load spectrum was further reduced to a constant load range. After a number of equivalent cycles, this load range results in the same equivalent accumulated damage as the spectrum of loads previously calculated through the rainflow counting algorithm. The short-term DELs were calculated following Equation (1):

$$\boldsymbol{S_0} = \left[\frac{\sum_i n_i * S_i^m}{n_{eq}}\right]^{\frac{1}{m}} \quad (1)$$

where n_{eq} is the equivalent number of cycles, S_i the different load ranges, n_i the corresponding cycle numbers and m is given by the slope of the Stress-Cycle (S-N) curve of the material used for the tower (DNV/Risø, 2002). In this case, it ~~was~~ is assumed an inverse slope~~that~~ $m = 3$ for the steel tower and $n_{eq} = 9.5064$ for the 10 min time series (equivalent to 10^7 cycles in 20 years). The DELs were then used as the dependent variable of the model.

2.2 Methods

The ~~main~~ methodology used ~~for the development of~~ in this paper is graphically described in Figure 1. First, the sensors which provide relevant information to model resultant fore-aft tower bending moments were selected (see Table 1). In the next step, the resulting records ~~was analysed~~were analyzed for missing data (e.g. zero values) and outliers. There were periods where the turbine was out of service or measurement failures with no registered data. Subsequently, affected records have been removed. The process of outlier detection in this study was not automated but done through visual inspection of the descriptive statistics calculated from the time series for each operational mode. To determine the relationship between the dependent and explanatory variables described previously, each of the 10 min files was summarized by estimating the following descriptive statistics for every explanatory variable: i) minimum value, ii) maximum value, iii) arithmetic mean, iv) range, v) mode, vi) standard deviation, and vii) variance.

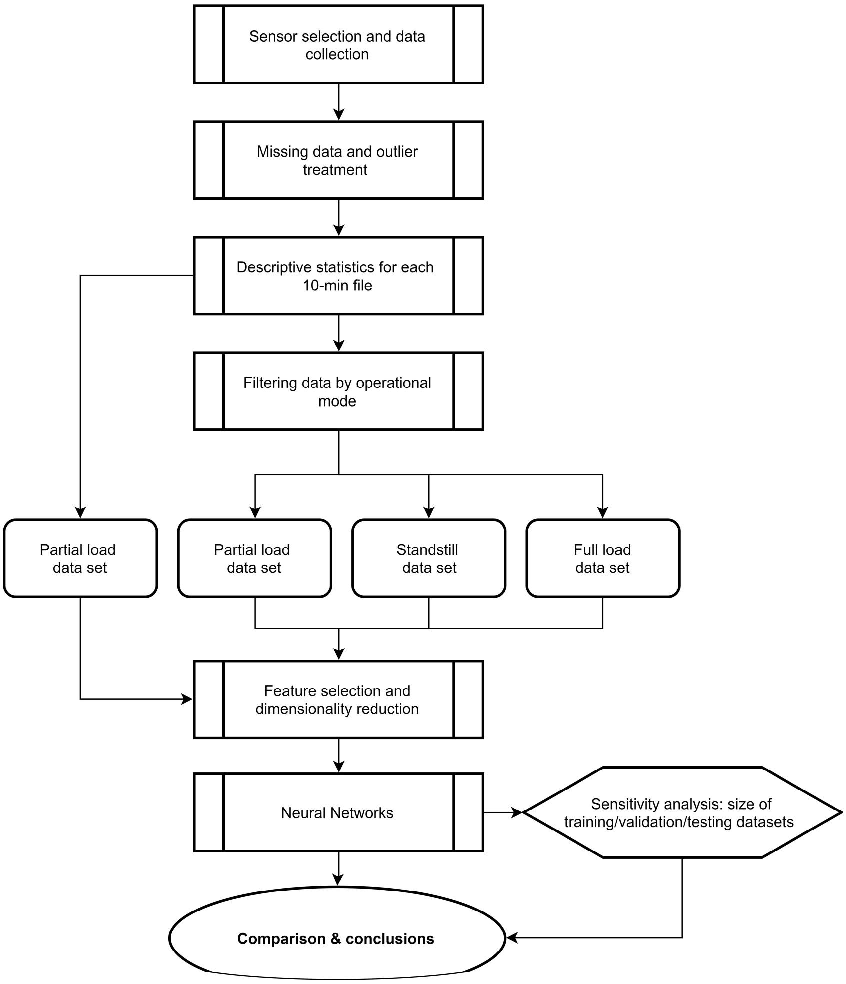


Figure 1 –~~:~~ Main methodology steps

In this way, the dataset was reduced to 63 features or explanatory variables. Excluding the time where no SCADA data was recorded, the total amount of data results in 36266 (77.1 %) observations. -This corresponds to a little over eight months of usefull information. ~~(corresponds to nearly 6044 hours).~~

Furthermore, ~~for the sensitivity analysis,~~ the data was filtered by operational modes, namely standstill, partial load, and full load. This was done by means of the feature "ACpow" which refers to a~~A~~ctive power output. In this sense, standstill corresponds to 10 min mean "ACpow" readings below or equal to 5 kW (0.25% of nominal power); partial load to readings higher than 5 kW and below or equal to 2000 kW (97.56 % of nominal power); and full load to readings above 2000 kW.

Research by Sharma and Saroha (Sharma and Saroha, 2015) concluded that a reduction of dimensions possibly leads to a better performance of the mining algorithms while maintaining a good accuracy, therefore, it is important to eliminate potential redundant data and select the variables with ~~more~~ most predictive power for the model. For this, ~~four~~ three different feature selection techniques and one dimension reduction technique~~s~~ were applied to the entire dataset and the datasets resulting from filtering the data by operational mode.

These techniques include Pearson correlation, stepwise regression, NCA, and PCA. Pearson correlation measures the linear correlation between two variables and maps the result to an interval between -1 and 1, where 0 indicates no linear relationship (Boslaugh and Watters, 2008). It can be calculated as per Eq. (2):

$$\boldsymbol{r_{XY}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2}$$

where n is the sample size, X_i and Y_i are the observations with index i and $\bar{X}$ represents the arithmetic mean of all ~~the~~ samples. A threshold value of 0.5 was set to define the level of correlation. In this sense, a correlation coefficient between 0 and 0.49 is weak and a correlation coefficient between 0.5 and 0.95 is strong. Correlation among all features of a particular sensor above 0.95 was considered as a redundant sensor and therefore eliminated for further analysis.~~Correlation coefficients above 0.95 may indicate redundancy in the dataset and lead to lower model accuracy.~~

Stepwise regression is an iterative method where features are added and removed from a multilinear model based on their statistical significance in the regression (Draper and Smith, 1998). The algorithm begins by constructing an initial model with one feature (forward selection) or all the features (backward selection) and continues adding or removing features by comparing the explanatory power of the larger or smaller models. At each step, the p-value of the corresponding F-statistic is estimated and compared to a threshold p-value to decide which features are included in or excluded from the model. P-value is used as a probability measure to identify if a particular feature is significant for the outcome of the model. If a p-value is larger than 0.05 the null hypothesis is true and the feature is selected for further modeling. The algorithm repeats this process until the added feature does not improve the model ~~anymore~~ or until all features that do not improve the explanatory power of the model are removed. This method is considered to be locally optimal, yet not globally optimal given that the selection of features included in the initial model is subjective and there is no guarantee that a different initial model will not lead to a better fit.

NCA is a non-parametric classification model used for metric learning and linear dimensionality reduction (Goldberger et al., 2005). It is based on a mode~~l~~ling technique known as k-Nearest Neighbours (k-NN), which is a supervised learning algorithm used for classification or regressions (Han and Kamber, 2006; Parsian, 2015). In its simplest form, the k-NN approach looks for the closest k = 1 observation to the query observation x_q within the training dataset by measuring the distances to the neighbouring data points and selecting the one that satisfies $\min_i$ distance (x_i, x_q). The output is then predicted by applying a function y = h(x) where h is the trained k-NN prediction function. In a multidimensional dataset, the k-NN approach requires to differentiate between the "relevance" of the explanatory variables for the intended output. For the learning process,~~For this purpose,~~ different weights can be assigned to the features of the model using the "Scales Euclidian Distance" estimation detailed in Eq. (3):

$$Distance(x_i, x_q) = \sqrt{a_1\left(x_i[1] - x_q[1]\right)^2 + \cdots + a_d\left(x_i[d] - x_q[d]\right)^2} \tag{3}$$

where x_i is a vector of input values, x_q is the query vector, a is the scaling number that defines the relevance of each explanatory value, and d the total number of ~~observations~~features. The weights are assigned randomly and then adjusted by solving a minimization problem (minimizing the prediction error). Other distance metrics can be used, namely Mahalanobis, Manhattan, rank-based, correlation-based, and Hamming (Hazewinkel, 1994).

Lastly, ~~principal component analysis (~~PCA~~)~~ is a statistical method to reduce the dimensions of a dataset that presumably contains a large number of irrelevant features while retaining the maximum information possible (Vidal et al., 2016). This is done by transforming the original set of multidimensional data into a new set referred to as components by means of eigenvectors and eigenvalues. A pair of eigenvector and eigenvalue indicate respectively the direction and how much variance is there in the data in that direction. The eigenvector with the highest eigenvalue is the first principal component. In this sense, the transformation allows ~~to reduce~~reducing the dimensions of the dataset to a few components with relatively low loss of information.

In this way, 16 neural networks (NN) were developed corresponding to four datasets (all operational modes, standstill, partial load, and full load) and ~~four~~ three feature selection techniques and one dimension reduction technique~~s~~. Each dataset is divided into training, validation, and testing subsets. 70 % of a dataset is randomly chosen and used by NN for training the model, 15 % are used to testing and 15 % for validation, i.e., this subset is used to adjust the model by means of the mean squared error (MSE). This adjustment stops when the MSE does not significantly improve. The validation subset is used as a measure to avoid overfitting the NN and generalize~~s the transfer function~~ the prediction model. ~~so that the~~After that, the model applies ~~is applicable~~ to new datasets. The test subset does not affect~~has no effect~~ on training or validation, it is only used to measure the performance of the trained NN.

The NN models used in this paper are trained with the Neural Network Toolbox from MATLAB (MathWorks, 2019). The standard settings consist of a two-layers feed-forward NN with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. The number of hidden neurons in the hidden layer are kept, as suggested by MATLAB, at 10 neurons. The Levenberg-Marquardt algorithm is selected as the training algorithm. No changes have been made to the standard configurations suggested by MATLAB.

The results from the 16 models were compared to derive conclusions about the relationship between operational data and tower ~~fatigue~~ loads acting on WTs.

Finally, the predictive capability of the model for continuous monitoring is tested. For this purpose, the neural network is trained using only the first 50 % of the data gathered during partial load. The prediction error is estimated to determine the accuracy of the model.~~Finally, a sensitivity analysis was developed by varying the sizes of datasets for training, validation, and testing in the partial load model to determine the required length of the data collection campaign necessary to achieve accurate results efficiently.~~

3 Results and discussion

This section describes the sensors identified by different methods as potential predictors of tower fatigue loads of the WT~~. Additionally, it assesses the outcomes of the models used to predict the fatigue loads. Finally, it~~ and presents the results of using a predictive model for continuous monitoring.~~of the sensitivity analysis performed that sought to determine the length of the data collection campaign required to appropriately predict the fatigue loads.~~

3.1 Feature selection and dimension reduction

Before building a model to predict the desired output, it is important to define which variables could act as predictors. The feature selection methods described in section 2.2 were applied to four different datasets: i) an eight months ~~complete one-year~~ dataset, ii) a full-load dataset, iii) a partial load dataset, and iv) a standstill dataset. The results of the feature selection methods are described below. For detailed information on selected features for each operational mode, the reader is referred to appendix A 1 to A 3.

3.1.1 Complete dataset: eight months ~~one-year~~ data

A Pearson correlation analysis was applied to the pre-selected features for predicting the DELs of the fore-aft bending moment of the tower finding that only ~~25~~ 27 of the ~~63~~ 56 features are correlated and used as independent variables in the model. The results indicate that the accelerations in both directions (i.e., x and y-axis) are highly correlated with the DELs. The standard deviation of the acceleration in the x-direction presents the highest correlation with a coefficient of 0.97, depicting an almost linear relationship between this feature and the dependent variable.~~. On the one side, the results indicate that most of the descriptive statistics of the acceleration meter in both directions (i.e. x and y-axis) are the most highly correlated features with the DELs. As a matter of fact, the standard deviation of the acceleration in the x-direction presents the highest correlation with a coefficient of 0.97, depicting an almost linear relationship between this feature and the dependent variable.~~ The maximum and mean ~~occurring~~ windspeed, and ~~both the average and maximum~~ power output are also highly correlated with the DELs. As an example, the relationship between the mean wind speed and the calculated DEL is graphically shown in Figure 2.

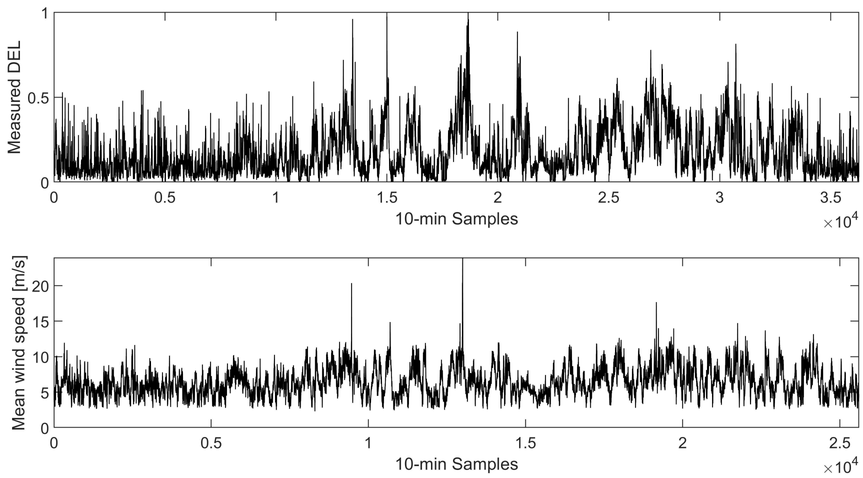


Figure 2 – Normalized DELs and mean wind speed

These results suggest that these features fluctuate together with the dependent variable, and they could potentially be used to build a model that can estimate the DELs of the fore-aft bending moments of the tower without installing strain gauge~~s~~ sensors. Furthermore~~On the other side~~, the results show that air density and relative wind direction, along with all their corresponding descriptive statistics, have a very low correlation with the DELs and should be, therefore, disregarded in the model based on this feature-selection technique. The mean wind direction, in particular, has a correlation coefficient close to zero, indicating an insignificant linear relationship with the DELs. ~~Since the DELs are calculated for the thrust loads, the highest observed value for this variable occurs when the turbine is directly facing the wind and the lowest when the wind is not affecting the thrust load directly.~~

Before using the features with the strongest correlation in a model, it is necessary to check for collinearity, i.e., correlation between independent variables. By observing the correlation between the explanatory variables, it can be determined which variables are highly correlated with each other and, therefore, should be excluded from the model to avoid collinearity issues. From this analysis, it was determined that rotational speed at the rotor should be excluded from the model and only rotational speed at the generator should be included given that these two features are a factor away from each other and, thus, may add bias to the model due to redundancy. This resulted in 56 features from initial 63. Many of the remaining variables are also highly correlated with each other, nevertheless, they add potentially valuable information to the model.

An alternative would be the use of a method such as PCA which could contribute to avoiding multicollinearity by transforming the data while maintaining the information contained in it.

After removing the rotational speed at the rotor and eliminating all variables with a correlation below 0.5, ~~there are 21 features~~12 feature are remaining after using PCA that could be included in the model. This data was transformed as explained in section 2.2 estimating the variance explained by each of the first components as seen in Figure 3. It can be observed that 99 % of the information contained in the features is now stored in the first ~~nine~~ 12 components. The remaining ~~13~~ 15 components explain less than 1 % of the cumulative variance. A model could be built using the first ~~nine~~ 12 components and the results should be almost as accurate as using the ~~21~~ 27 features selected after the correlation analysis. The biggest disadvantage with this method is that given the transformation of the data, it is no longer possible to interpret it. The results, nevertheless, remain interpretable and are free of the influence of multicollinearity.

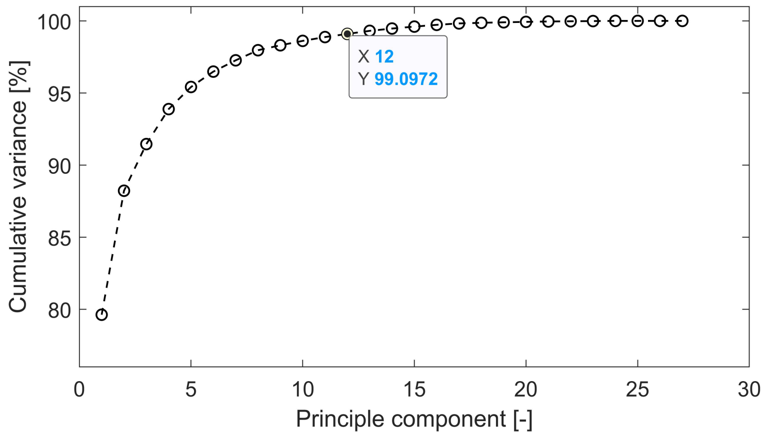


Figure 3 – ~~:~~ Cumulative variance explained by principal components

Alternatively, an interactive stepwise regression was built using the pre-selected ~~63~~ 56 features. Different combinations of features were tested to identify those that should not be included in the model given that they do not contribute to the predictive power or they result in an increase in the error of the model. The features with a p-value above 0.1 should be omitted from the model. ~~Unlike the Pearson correlation analysis, this method avoids multicollinearity among the features.~~ The results suggest excluding a total of ~~26~~ 30 variables from the regression model. Among these can be found the minimum, maximum, mean, and range of the rotational speed at the generator, most descriptive statistics of air density, with the exception of the standard deviation, and range, mode, and standard deviation of the acceleration in y-direction.

It is important to highlight that the variable that represents the range of the acceleration sensor in the x-direction was identified as statistically insignificant despite its high correlation with DELs. As mentioned earlier, the possible models explored with the stepwise regression are limited. The algorithm builds different models from the ~~63~~ 56 features depending on the order in which these features are added to (in the case of forward selection) or removed from (in the case of backward elimination) the models. In this sense, the range of the sensor “acc_x” and the variance of “acc_x”, which are correlated with a factor of 0.90, could be considered mutually exclusive. The decision as to which of these variables to include in the model would depend solely on which variable is added or removed first in the stepwise regression. In this case, the algorithm suggests excluding the range of “acc_x”, a highly correlated feature, based on the search for the local minimum instead of evaluating all combinations. Ultimately, this method identified ~~37~~ 33 features as statistically significant and, thus, these should be included in the model.

The last feature selection method, NCA, was applied as well to the initial 63 features dataset. ~~Ten~~ 13 features were identified by this method as relevant for the prediction of DELs, a significantly smaller number than those selected by applying the correlation analysis and stepwise regression.

To summarize, mean values and standard deviations are the descriptive statistics that can best describe the data according to the three feature selection methods applied. Features selected by all three methods include wind speed, acceleration, and power output.

3.1.2 Data filtered by operational modes

The dataset was divided by operational modes resulting in 7825 10-min samples (21.6 %) of the data corresponding to standstill, 2837 10-min samples (7.8 %) to full load, and 25604 10-min samples (70.6 %) to partial load. Each dataset contains 56 features and the corresponding DELs. ~~Unlike with the entire dataset, the rotational speed at the rotor was removed due to redundancy with the rotational speed at the generator.~~

The first feature selection method used in these new datasets is ~~the~~ again the Pearson correlation analysis. The results show that most of the descriptive statistics for wind speed and acceleration are highly correlated with the DELs in all operational modes. The first differences appear in the generator speed. As expected, the generator speed is not relevant during stand~~-~~still since the rotor is not moving or is only idling. The mean wind speed is not as relevant in full load as it is in partial load. During full load, the rotational speed is around a specified number and must be kept as stable as possible. Therefore, the mean rotational speed does not change significantly during full load. During partial load, the mean rotational speed is within a higher range, therefore it has a higher correlation with the corresponding DELs.

During full load, the standard deviation and variance of the rotational speed are highly correlated with the DELs. The standard deviation explains how the values differ from the mean, thus conclusions about the dynamics of the turbine can be derived based on these spreads. For example, a large deviation in the rotational speed around the mean during full load has a significant effect ~~of~~ on the tower movement. The features of the pitch angle are correlated with the output ~~with~~ of a factor greater than 0.5 with the output. This correlation is only significant during full load. The pitch angle is held at the most efficient lift-to-drag ratio during the partial load and, therefore, not many variations can be observed during standstill and partial load. During full load, the turbine pitches constantly to keep the rotational speed nearly constant. For each operational mode, PCA was performed to account for potential collinearity in the feature-set. This was done consistently with an explained variance of 99 % remaining.

The second feature~~s~~ selection method applied is ~~a~~ stepwise regression. The results are not consistent with the correlation analysis. Air density and wind direction ~~had no correlation~~did not correlate with the DELs, however, they were chosen by the stepwise regression during standstill and partial load as potential predictors. Also, the pitch was chosen as a significant variable during standstill, even though the turbine is not pitching. In general, the modeler needs to be careful when interpreting the results from a stepwise regression as described in section 2.2.

~~Lastly,~~ NCA was applied to the three datasets. Examining the results, one significant difference to the correlation analysis is that the wind direction was identified as significant during standstill and partial load by the NCA, whereas the correlation analysis showed no correlation of these features with the output during any operational mode. Furthermore, the range of the pitch angle was identified as relevant during partial load, which was not the case in the correlation analysis. Mean acceleration in the x-direction and the standard deviation were the only two features identified as significant by the NCA in all three operational modes.

3.2 Modeling fatigue loads

Once the features with predictive power have been identified for the different datasets, NN are built to evaluate the predictions. The outcomes of these models are described hereafter.

3.2.1 ~~One-year~~Eight month data with all operational modes

The first analysis is conducted on eight month data without filtering by operational modes. To illustrate the results, features used for training the NN are selected by using the correlation analysis and can be examined in appendix A 1. The data is randomly split into training, testing and validation sets. The regression model of the NN in Figure 4 shows a similar R-value in all regressions indicating that there is no overfitting in the model.

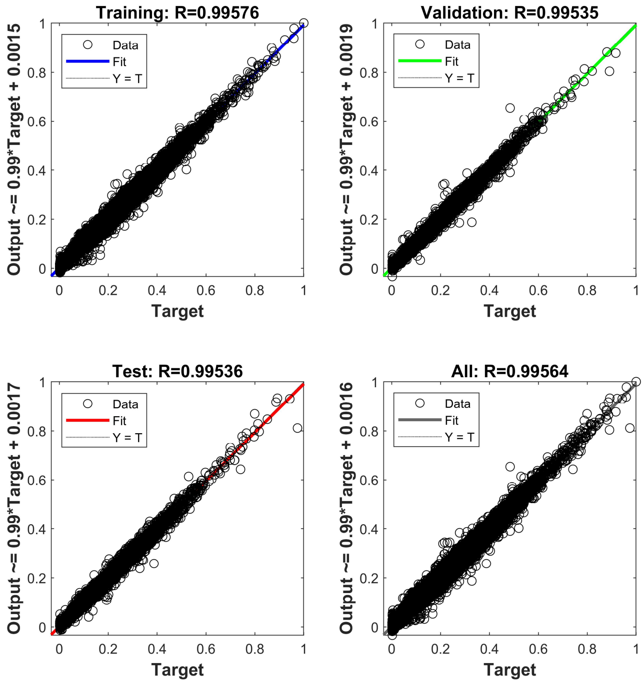


Figure 4 – Linear regressions between the normalized neural networks prediction and the measured DELs

The R-value for the complete data-set is 0.99564 which can be confirmed by observing the top plot in Figure 4 where the predicted DELs overlaps with the measured DELs with a mean prediction of 2.22 % (see Table 2 below). Nevertheless, the prediction error can be as high as 685.79 %. Values close to zero can have a significant impact in terms of the mean error in percent due to a high ratio of prediction and measured DELs. Figure 5 shows the normalized predicted vs. measured DELs and the corresponding prediction error.

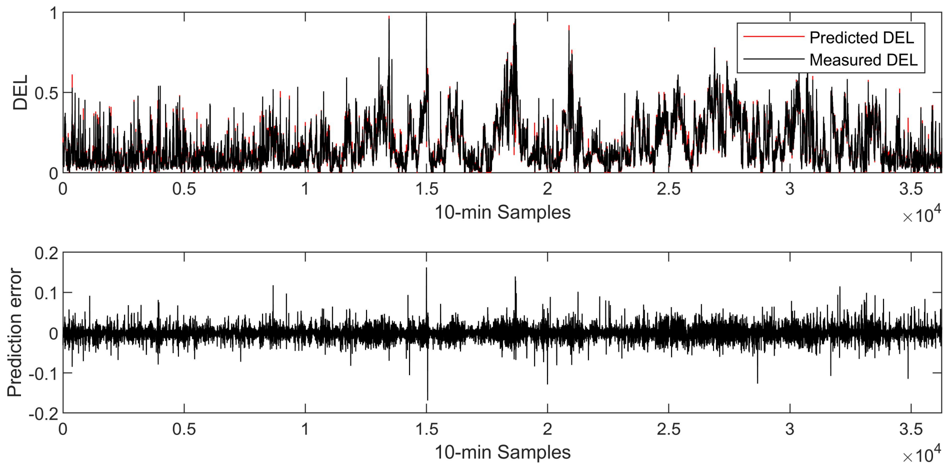


Figure 5 – NN prediction: The top plot presents the normalized predicted and the measured DELs by NN. The button plot is the prediction error.

Plotting the error against the wind speed for all operational modes in Figure 6, it can be concluded that the mean prediction error is significantly higher at low wind speeds. If the wind speed is below 5 m s^{-2} the WT is in a standstill.

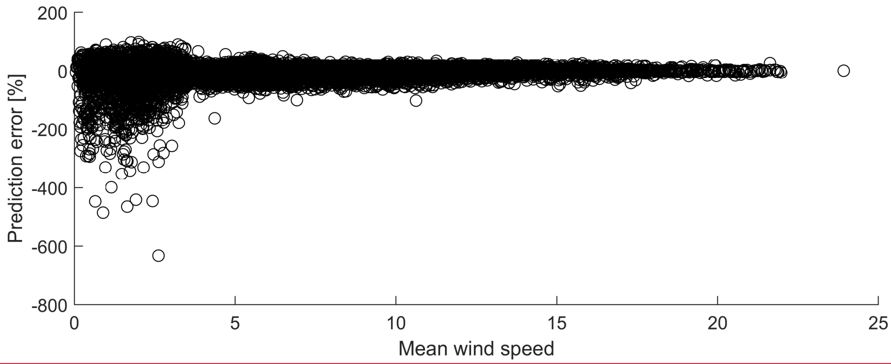


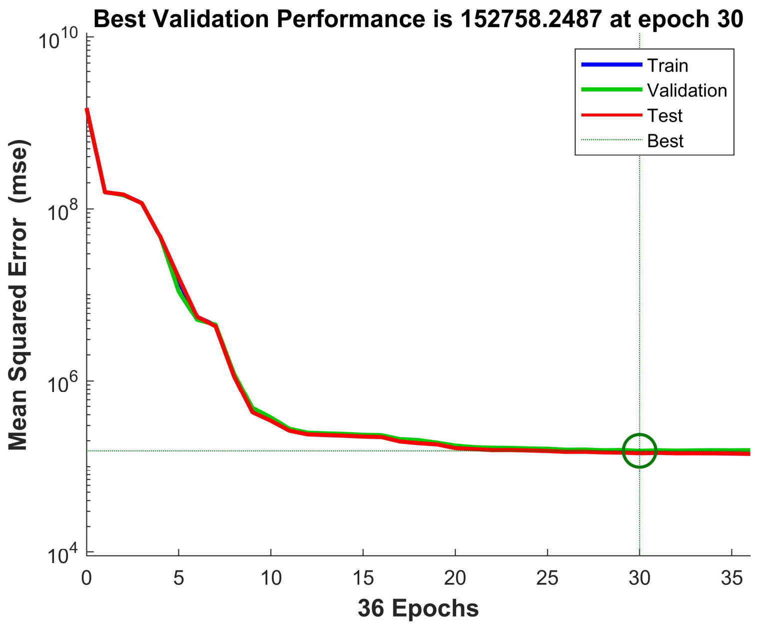
Figure 6 – Wind speed against prediction error

Comparing the results in Table 2, it can be seen that the model using features selected by NCA results in the lowest mean error. Overall, these results are significantly higher than those obtained by Vera-Tudela and Kühn (2014). This can be explained by the high prediction error during low wind speeds as seen in Figure 6. Additionally, the results indicate that it is possible to significantly reduce the number of features used in the model by applying NCA while maintaining a low prediction error.

Table 2 – Summary of results from Neural Networks for the complete one-year considering all operational modes

N°	Feature subset	No. of features (% of total)	R			Mean error [%]	Std dev [%]	Max abs error [%]	Mean abs. Error [kNm]
			Training	Validation	Test				
1	Correlation	27	0.99576	0.99535	0.99536	2.22	22.85	685.79	237
2	Correlation & PCA	12	0.99393	0.99400	0.99365	2.94	18.93	410.33	276
3	Stepwise	33	0.99555	0.99523	0.99476	2.24	7.23	671.48	224
4	NCA	13	0.99581	0.99593	0.99568	2.07	26.09	525.20	228

~~The NN stopped its training at epoch (iteration) 30 with the best performance at a mean squared error of 152758 kNm (see Figure 3). The performance plot shows how the mean square error drops rapidly as the NN learns. Six more iterations were made to confirm that the value does not change significantly.~~



~~Figure 3 – Neural Networks performance validation~~

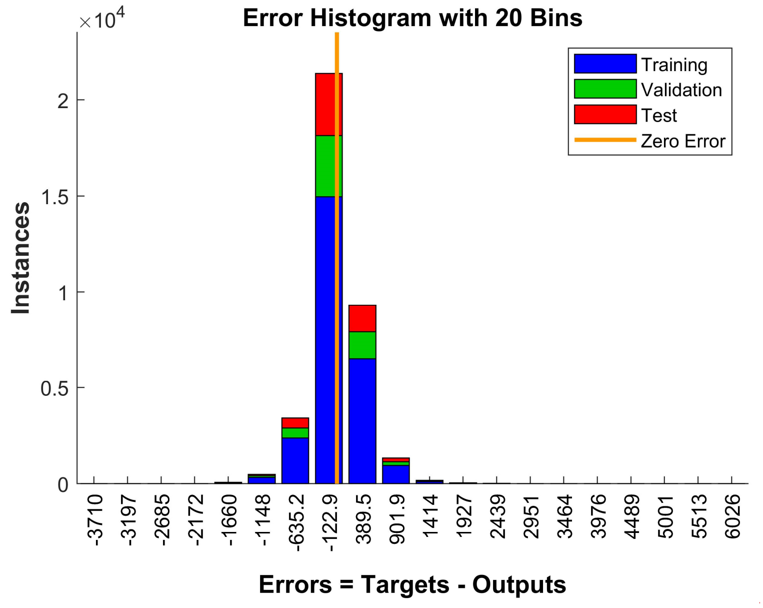
~~From the error histogram in Figure 4, it can be observed that most errors lay in the interval from -1000 to 1000. Approximately 23567 of the data points (65 % of the total) have an error between -500 and 500. Data points close to -3500 kNm and 4500 kNm differ significantly from the mean and occur less frequently, which could indicate the presence of outliers in the data used for modeling. These potential outliers can also be observed in the regression plot in Figure 5, where some values are clearly distanced from the regression line. The regression plot shows the calculated DELs from the strain gauges measurements with respect to the trained DELs obtained from the NN. The correlation "R" is significantly high for training, validation, and~~

~~test data. The R on the testing data is 0.99486 and higher in the cases of training and validation data, which indicates a nearly perfect positive linear correspondence. All values need to lay on the 45-degree line to have a perfect fit. This can only happen if the network output equals the target values.~~

~~The calculation for mean error in percent following Eq. (4):~~

$$Error\ [\%] = \frac{Target_{value} - Modeled_{value}}{Target_{value}} * 100 \qquad (4)$$

~~Applying the Eq. (4) and dividing by the number of observations, a mean error of 9.45 % and a maximum error of 526 % can be calculated. The maximum error indicates a possible outlier in the dataset that confirms the errors observed in the histogram.~~

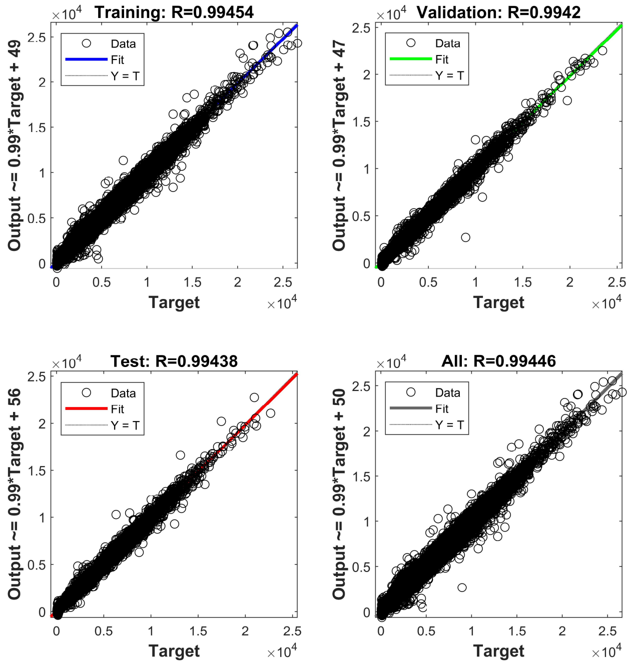


~~Figure 4 – Neural Networks error histogram~~

~~Table 2 summarizes the results of neural networks built from the different predictor sets.~~

~~Table 2 - Summary of results from Neural Networks for the complete one-year dataset and data subsets~~

~~N°~~	~~Subset~~	~~No. of features~~	~~R~~			~~Iterations (epoch)~~	~~Max Error [%]~~	~~Mean abs. Error [%]~~	~~Mean abs. Error [kNm]~~
			~~Training~~	~~Validation~~	~~Test~~				
~~1~~	~~Correlation~~	~~31~~	~~0.99584~~	~~0.99554~~	~~0.99536~~	~~30~~	~~526~~	~~9.45~~	~~237~~
~~2~~	~~Correlation & PCA~~	~~9~~	~~0.99437~~	~~0.99379~~	~~0.99405~~	~~83~~	~~444~~	~~10.85~~	~~276~~
~~3~~	~~Stepwise~~	~~38~~	~~0.99632~~	~~0.99569~~	~~0.99556~~	~~28~~	~~1134~~	~~10.38~~	~~224~~
~~4~~	~~NCA~~	~~13~~	~~0.99581~~	~~0.99593~~	~~0.99568~~	~~49~~	~~980~~	~~10.49~~	~~228~~



~~Figure 5 – Linear regressions between the Neural Networks prediction and the DELs~~

~~Comparing the different subsets trained with the same settings in the NN, it can observe that the correlation "R" value is high in all cases. More significant differences can be observed in the maximum and mean error. It is presumed that outliers lead to a 1000 % maximum error in the case of the stepwise regression subset and maximum errors above 400 % in the remaining models. The mean absolute error of approximately 10 % shows that most values do not present high errors such as 1000 %, therefore, it can be concluded that the general performance of the model is acceptable and can be improved through outlier treatment and data smoothening.~~

3.2.2 Data filtered by operational modes

This section explores the performance of the NN when built with data subsets from different operational modes. It can be observed that the mean error in percent is significantly high in standstill compared to other operational modes as shown in Table 3. Nevertheless, the mean absolute error in kNm is the lowest. The high maximum error observed previously when using the complete ~~one year~~eight month dataset (i.e., when using "all operational modes") for the different training sets could be explained by the poor predictive power of the data from the standstill mode. When the NN was built using filtered data for partial and full load, the errors of the predictions decreased significantly. Thus, it can be concluded that the data from the standstill mode adds uncertainty to the model. This can be observed in Figure 7.~~Thus, the data from the standstill mode adds noise and reduces the accuracy of the model.~~

Presumably, the small variations observed in the readings from the sensors during standstill do not provide enough information to predict the DELs. This is consistent with the R-values of the model during standstill mode. These values are the lowest among the different operational modes. A more detailed look at this case would be necessary to derive valuable insight.

(a) (b) (c)

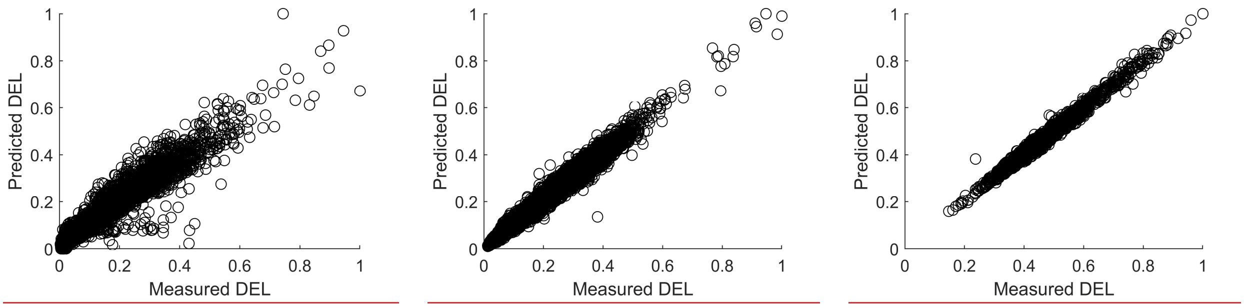


Figure 7 – Normalized measured vs. predicted DELs. Figures correspond to a model built with features from the correlation analysis. Subfigure (a): Standstill. Subfigure (b): Partial load. Subfigure (c): Full load.

Moreover, Table 3 shows that the partial and full-load models constructed using smaller sets of features derived from the application of methods such as PCA or NCA have approximately the same predictive power as those models constructed using larger sets of features derived from applying methods such as stepwise regression or correlation analysis. This can be observed in the comparison of the measures of goodness-of-fit (i.e. R-values) among the models. Thus, the application of feature-selection and dimension-reduction methods can be considered a good practice. The results from the full load model can be compared to the existing work from Vera-Tudela and Kühn (2014) where the mean error is below 0.22 for all feature-sets. The maximum absolute error and standard deviation of the error also confirm the results. The accuracy of the results from the partial load model is slightly worse for all features-sets.

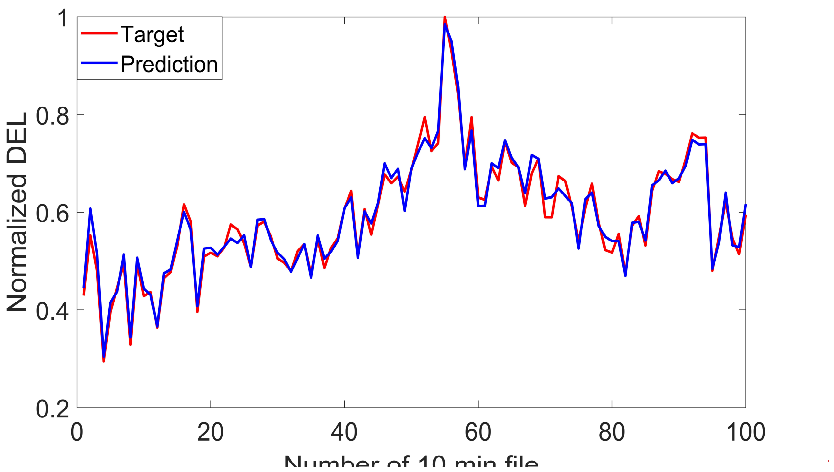
Table 3 – Summary of results from Neural Networks for different operational modes

N°	*Subset*	*No. of features*	*R*			*Mean error [%]*	*Std dev [%]*	*Max abs error [%]*	*Mean abs. Error [kNm]*
			Training	*Valida-tion*	*Test*				
Standstill									
1.1	Correlation	18	0.96884	0.96598	0.96962	9.89	37.46	472.83	189
1.2	Correlation & PCA	7	0.96211	0.96776	0.95255	14.13	40.01	432.56	199
1.3	Stepwise	35	0.98781	0.98352	0.97053	7.21	46.51	828.72	138
1.4	NCA	16	0.98369	0.98163	0.97891	6.60	39.77	505.83	145
Partial Load									

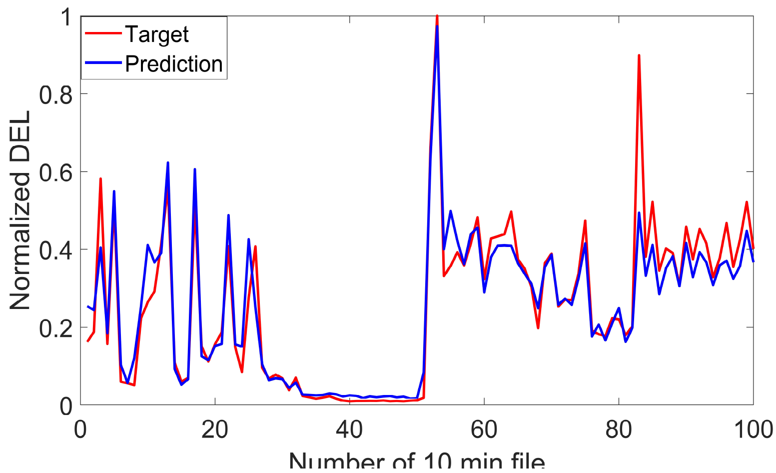
2.1	Correlation	28	0.99322	0.99228	0.99212	0.70	8.77	82.34	242
2.2	Correlation & PCA	12	0.99045	0.99034	0.99040	0.69	9.74	89.06	256
2.3	Stepwise	37	0.99330	0.99295	0.99217	1.21	9.03	71.16	239
2.4	NCA	11	0.99282	0.99236	0.99218	0.72	9.33	79.02	240
				Full Load					
3.1	Correlation	28	0.99186	0.98858	0.98389	0.06	3.07	56.28	276
3.2	Correlation & PCA	12	0.98759	0.98568	0.98636	0.11	3.58	53.35	290
3.3	Stepwise	23	0.99175	0.99011	0.98953	0.07	2.79	16.29	272
3.4	NCA	8	0.99048	0.98847	0.99084	0.07	3.09	54.85	273

Nevertheless, it is important to mention that to apply PCA the complete feature-set is needed to transform all the information in the first few components. This is not the case when applying NCA where the most relevant features are directly identified. ~~Figure 6 shows that DELs estimations during standstill were significantly lower compared to those during full load (DELs in Figure 6 (b) are around three times lower than during full load Figure 6 (a)). Therefore, an error of over 20 % in standstill mode is equivalent to a 7 % error in full load. This, nevertheless, does not justify the 20 % error seen during standstill mode. The NN performs worse when using data from the standstill mode than when using data from the other operational modes. Presumably, the small variations observed in the readings from the sensors during standstill do not provide sufficient information to predict the DEL. This is consistent with the R values of the model during standstill mode. These values are the lowest among the different operational modes. A more detailed look at this case would be necessary to derive valuable insight.~~

~~(a)~~



~~(b)~~



~~**Figure 6 – Comparison of target vs. prediction in full load (a) and standstill (b) mode**~~

~~Moreover, Table 3 shows that those models constructed using smaller sets of features derived from the application of methods such as PCA or NCA have approximately the same predictive power as those models constructed using larger sets of features derived from applying methods such as stepwise regression or correlation analysis. This can be observed from the comparison of the measures of goodness-of-fit (i.e. R^2-values) among the models. It is important to highlight that this observation is valid for all operating modes. Thus, the application of feature-selection and dimension-reduction methods can be considered a good practice.~~

~~**Table 3 – Summary of results from Neural Networks for different operational modes and data subsets**~~

N°	Subset	No. of features	R			Iterations (epoch)	Max Error [%]	Mean abs. Error [%]	Mean abs. Error [kNm]	R^2
			Training	Valida-tion	Test					
					Standstill					
1.1	Correlation	18	0.97029	0.95458	0.96413	17	477	26.54	189	0.935
1.2	Correlation & PCA	7	0.96498	0.9623	0.95612	26	350	21.80	199	0.928
1.3	Stepwise	35	0.98551	0.9809	0.97639	34	1159	22.63	138	0.967
1.4	NCA	16	0.97957	0.98052	0.97744	21	600	24.64	145	0.959
					Partial Load					
2.1	Correlation	28	0.99297	0.99233	0.99201	19	63	6.65	242	0.985
2.2	Correlation & PCA	12	0.99169	0.99046	0.99108	43	84	6.97	256	0.983
2.3	Stepwise	28	0.9932	0.99211	0.99202	18	69	6.57	239	0.986
2.4	NCA	11	0.99277	0.99236	0.99216	38	82	6.59	240	0.985
					Full Load					
3.1	Correlation	28	0.99124	0.98997	0.98899	13	52	2.27	276	0.981

~~3.2~~	~~Correlation & PCA~~	~~12~~	~~0.99021~~	~~0.98793~~	~~0.98775~~	~~25~~	~~55~~	~~2.37~~	~~290~~	~~0.979~~
~~3.3~~	~~Stepwise~~	~~23~~	~~0.99169~~	~~0.99128~~	~~0.98713~~	~~16~~	~~58~~	~~2.24~~	~~272~~	~~0.982~~
~~3.4~~	~~NCA~~	~~11~~	~~0.99208~~	~~0.98754~~	~~0.98371~~	~~26~~	~~30~~	~~2.24~~	~~273~~	~~0.980~~

~~3.3~~ Continuous monitoring with a predictive model ~~Sensitivity analysis~~

In this section, the results of using a predictive model for continuous monitoring are presented. The aim is to identify if and how the errors in the outcomes of the model vary when using only the first 50 % of data gathered. The model is tested by predicting the DELs corresponding to the remaining share of the data. The majority (i.e., 70.6 %) of all data gathered corresponds to the partial load mode, therefore this subset was selected for this analysis. As in the previous analysis, four models are built using the feature-sets from the correlation analysis, correlation and PCA, stepwise regression, and NCA.
Figure 8 shows the prediction error from the model using the feature-set from the correlation analysis. It can be seen that the mean absolute prediction error from the model trained using the first 50 % of the data gathered is 211 kNm (see Table 4). This value is lower than the mean absolute error from the model trained with the remaining data which yields 244 kNm in Table 5. Nonetheless, the contrary is true for the mean error (in percentage). This variation can be explained by the same relationship observed previously in Figure 6 and Figure 7 where the prediction error decreases at high wind speeds. As can be seen in Figure 9, the average mean wind speed is higher in the second half of the partial load dataset, which was used for testing.

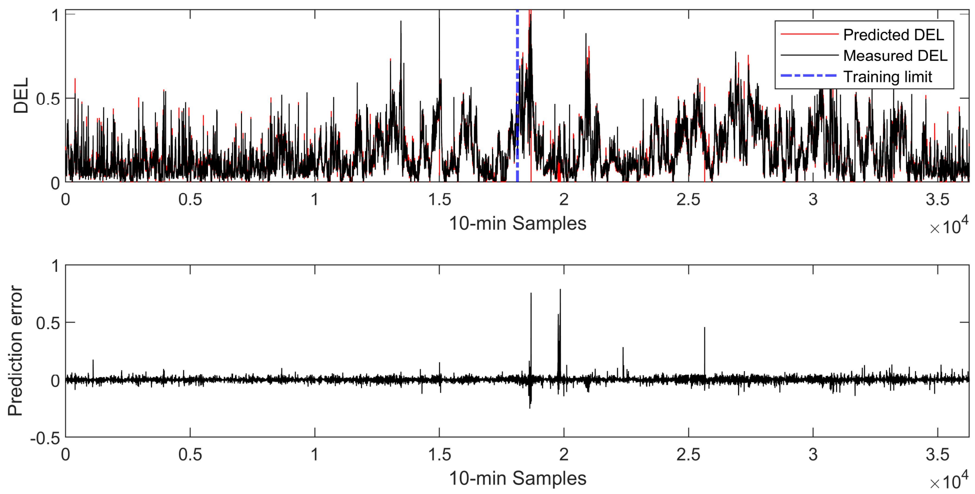


Figure 8 – Comparison of normalized predicted and measured DELs and the corresponding prediction error for the model using the feature-set from the correlation analysis.

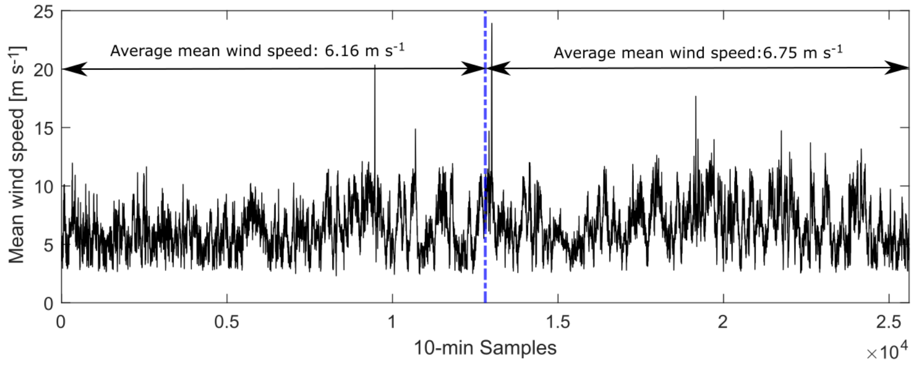


Figure 9 – **Mean wind speed during partial load**

The same behavior is observed in the remaining models. Overall, the mean error is lower in the results from the models trained using the second half of the datasets as can be seen when comparing Table 4 and Table 5.

Table 4 – Summary of results of the partial load model trained with the first 50 % of the data

N°	Feature subset	No. of features (% of total)	R Training	R Validation	R Test	Mean error [%]	Std dev [%]	Max abs error [%]	Mean abs. Error [kNm]
1	Correlation	27 (48%)	0.99276	0.99011	0.99182	0.92	9.60	76.44	211
2	Correlation & PCA	9 (16%)	0.99393	0.99400	0.99365	1.44	11.03	94.23	229
3	Stepwise	38 (68%)	0.99555	0.99523	0.99476	2.45	11.24	75.56	231
4	NCA	13 (23%)	0.99581	0.99593	0.99568	1.32	10.21	75.09	213

Table 5 – Results from using the trained model to predict the remaining 50 % of the data

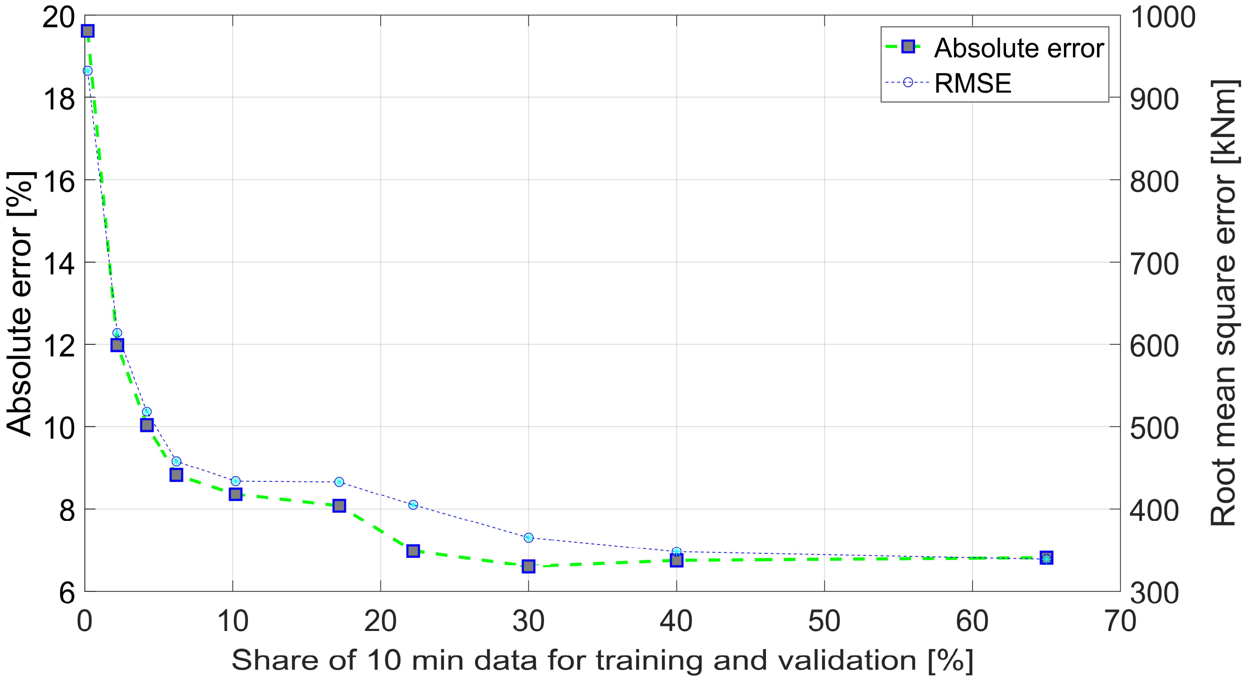
N°	Feature subset	No. of features (% of total)	Mean error [%]	Std dev [%]	Max abs error [%]	Mean abs. Error [kNm]
1	Correlation	27 (48%)	0.19	8.98	60.86	244
2	Correlation & PCA	9 (16%)	0.07	9.54	83.67	229
3	Stepwise	38 (68%)	0.84	9.55	88.64	248

4	NCA	13 (23%)	0.46	8.81	96.51	234

~~**In this section, the results of the sensitivity analysis are presented. The aim is to identify if and how the error in the outcomes of the model varied by reducing the size of the shares of observations used for training and validation of the NN. Since the majority (70.6 %) of the data gathered corresponds to partial load mode, this subset was used for the sensitivity analysis.**~~

~~The results from section 3.2.2 indicated that the errors did not differ significantly among the different sets of predictors used to predict the DELs during partial load, therefore the set of predictors obtained from the stepwise regression application were used for this sensitivity analysis.~~

~~Figure 7 shows that both the absolute error share and the root mean squared error decrease as the share of data used for training and validation increases. Nevertheless, this behavior stabilizes when using approximately 40 % of the data for training and validation. There is, then, a cost-reduction potential. Instead of investing time and effort in collecting large amounts of data over long periods of time, the same results can be obtained with a smaller dataset. In this case, 10241 observations were sufficient to yield an error that does not decrease significantly when increasing the share of training and validation data in the NN.~~



~~Figure 7 – Evaluation of dataset size for training / validation / testing a six month partial load predictive model~~

4 Conclusions

This paper used available SCADA data as well as strain gauge~~s~~ measurements from a research WT to develop a predictive model to estimate the DELs of the fore-aft bending moments of a WT tower. The dataset included a period of over eight months of~~6044 hours~~ useful data. Different feature selection methods and a dimension reduction technique were applied to choose the sensors with the strongest predictive power. The data were then inputted into a feedforward neural network. The methodology and data used reproduces and enhances the approaches of similar studies in the field of SHM.

The results indicate that using all data and applying neighborhood component analysis for feature selection yields an interpretable and low dimensional feature-set while maintaining high accuracy. ~~the most conservative model regarding the number of features and the most accurate outcomes with the lowest mean absolute error.~~ Additionally, dimension reduction techniques such as principal component analysis can contribute to a more parsimonious model reducing the number of features needed, however compromising the interpretability of the inputs given the transformation of the data.

The results were significantly better, i.e., yielded lower mean absolute errors, when the dataset was divided by operational modes. Particularly, the models were significantly more accurate when analyzing the operation of the turbine at full load and partial load. The outcome of the model using signals from when the turbine was standing still was rather inaccurate with mean ~~absolute~~ errors ~~approximately four times higher~~ ranging from 6 % to 14 %.~~than those from the partial load and approximately~~

~~10 times higher than those from full load.~~ In partial load the errors vary between 0.69 % to 1.21 % and in full load between 0.07 % to 0.11 %. It can be concluded that the performance of NN is influenced by the operational mode of the WT.

Finally, a model was built for continuous monitoring. For this, the first 50 % of partial load data was used for training and shows stable results in terms of prediction accuracy for the remaining data. All feature selection techniques showed similar results when predicting DELs for continuous monitoring. The feature set resulting from the application of correlation analysis and PCA yielded the lowest mean error, yet the second largest standard deviation for these errors. Since the results are not significantly different for each feature selection technique, the use of NCA is preffered due to its interpretable and low dimention of features.~~a sensitivity analysis performed in the partial load model determined that a data collection campaign of 1706 hours would suffice to build a neural network that provides outcomes with mean absolute errors equivalent to that of a neural network using data from 3627 hours.~~

This paper examined data from only one WT~~turbine~~. To be able to generalize the results obtained from this study, the NN model requires validation with data collected from a different wind turbine with the same specifications.~~To be able to generalize the results obtained from this study, the NN need to be trained with information from different turbines and data collected during different operating conditions.~~ By doing this, it will be possible to determine the relationship between SCADA data and fatigue loads with more precision, thereby eliminating the need to install expensive gauge sensors to estimate these loads and contributing to more efficient structural health monitoring methods.

Furthermore, the methodology developed during this study could be further tested by means of an analytical aero-elastic ~~analytical~~ model. Such a model would provide larger datasets from standstill and full load to test the predictive capabilities for continuous monitoring ~~datasets to train the NN more precisely~~ without the significant costs that this would imply if done empirically. The results of the NN trained with information from the aero-elastic model can be compared to the results presented in this paper to derive conclusions on the reliability and accuracy of this methodology.

Finally, the results could benefit from exploring alternative machine learning algorithms such as support vector machine and k-Nearest Neighbors.

5 Appendix A

A 1 Correlation Analysis by operational modes
Chosen values are highlighted in red for the correlation analysis.

Feature	***Still Stand***	***Partial Load***	***Full Load***	***All Modes***
acc_x_max	*0.92*	*0.92*	*0.88*	*0.96*
acc_x_mean	*0.91*	*0.93*	*0.96*	*0.96*
acc_x_range	*0.92*	*0.92*	*0.88*	*0.96*
acc_x_mode	*0.53*	*0.08*	*0.25*	*0.19*

Feature	Still Stand	Partial Load	Full Load	All Modes
acc_x_std	0.93	0.94	0.97	0.97
acc_x_var	0.77	0.85	0.94	0.86
acc_y_min	0.04	0.04	0.03	-0.08
acc_y_max	0.81	0.90	0.82	0.86
acc_y_mean	0.80	0.88	0.81	0.85
acc_y_range	0.81	0.90	0.82	0.86
acc_y_mode	0.35	0.14	0.00	0.06
acc_y_std	0.80	0.90	0.83	0.85
acc_y_var	0.70	0.77	0.82	0.75
v_wind_min	0.65	0.53	0.59	0.64
v_wind_max	0.76	0.90	0.87	0.80
v_wind_mean	0.74	0.83	0.84	0.74
v_wind_range	0.75	0.92	0.77	0.79
v_wind_mode	0.72	0.80	0.78	0.71
v_wind_std	0.73	0.94	0.81	0.77
v_wind_var	0.70	0.90	0.77	0.71
v_dir_min	0.12	0.08	0.12	0.16
v_dir_max	-0.13	-0.05	0.12	-0.16
v_dir_mean	-0.02	0.05	0.03	0.00
v_dir_range	-0.15	-0.07	0.01	-0.19
v_dir_mode	0.02	0.00	0.00	-0.01
v_dir_std	-0.15	0.07	0.11	-0.11
v_dir_var	-0.11	0.04	0.10	-0.08
omega_gen_min	-0.08	0.55	-0.79	0.64
omega_gen_max	0.19	0.80	0.81	0.68
omega_gen_mean	0.04	0.71	0.08	0.67
omega_gen_range	0.38	0.37	0.88	0.28
omega_gen_mode	0.02	0.65	-0.05	0.66
omega_gen_std	0.33	0.29	0.94	0.19
omega_gen_var	0.30	0.22	0.93	0.13
air_density_min	-0.10	0.22	0.05	0.23
air_density_max	-0.11	0.22	0.06	0.23
air_density_mean	-0.10	0.22	0.05	0.23
air_density_range	-0.02	0.01	0.08	-0.07
air_density_mode	-0.10	0.22	0.05	0.23
air_density_std	-0.02	0.00	0.09	-0.07
air_density_var	-0.02	0.02	0.07	-0.03

Feature	Still Stand	Partial Load	Full Load	All Modes
pitch_min	0.27	0.03	0.71	-0.35
pitch_max	0.41	0.31	0.87	-0.25
pitch_mean	0.35	0.21	0.81	-0.31
pitch_range	0.30	0.31	0.55	0.37
pitch_mode	0.36	0.12	0.66	-0.32
pitch_std	0.30	0.24	0.32	0.25
pitch_var	0.26	0.15	0.28	0.09
ACpow_min	-0.15	0.64	-0.05	0.82
ACpow_max	0.20	0.89	0.83	0.89
ACpow_mean	0.04	0.81	0.29	0.88
ACpow_range	0.21	0.92	0.17	0.70
ACpow_mode	0.00	0.75	-0.15	0.85
ACpow_std	0.21	0.88	-0.07	0.59
ACpow_var	0.20	0.70	-0.12	0.45

A 2 Stepwise Regression results for different operation modes
Chosen values are marked with a "x".

Feature	Still Stand	Partial Load	Full Load	All Modes
acc_x_min		x		x
acc_x_max	x	x	x	x
acc_x_mean	x	x	x	x
acc_x_range				
acc_x_mode	x	x	x	x
acc_x_std	x	x	x	x
acc_x_var	x	x	x	x
acc_y_min				
acc_y_max	x			x
acc_y_mean	x	x	x	x
acc_y_range	x			
acc_y_mode			x	
acc_y_std		x		
acc_y_var	x	x		x
v_wind_min		x	x	x
v_wind_max				
v_wind_mean	x	x	x	x
v_wind_range		x		
v_wind_mode	x		x	x
v_wind_std	x	x		x
v_wind_var	x	x	x	
v_dir_min	x			
v_dir_max	x	x		
v_dir_mean	x	x		x
v_dir_range				x
v_dir_mode				
v_dir_std	x	x	x	x
v_dir_var	x	x		x
omega_gen_min	x	x	x	
omega_gen_max		x		
omega_gen_mean	x	x		
omega_gen_range	x			
omega_gen_mode	x	x		x
omega_gen_std	x		x	x
omega_gen_var	x	x	x	x

Feature	***Still Stand***	***Partial Load***	***Full Load***	***All Modes***
air_density_min			*x*	
air_density_max				
air_density_mean				
air_density_range	*x*			
air_density_mode			*x*	
air_density_std	*x*	*x*		*x*
air_density_var	*x*	*x*		
pitch_min	*x*	*x*	*x*	*x*
pitch_max		*x*		*x*
pitch_mean		*x*	*x*	*x*
pitch_range				
pitch_mode	*x*	*x*	*x*	*x*
pitch_std		*x*	*x*	*x*
pitch_var	*x*	*x*		*x*
ACpow_min		*x*		*x*
ACpow_max	*x*	*x*	*x*	*x*
ACpow_mean	*x*	*x*	*x*	*x*
ACpow_range				
ACpow_mode	*x*	*x*		*x*
ACpow_std	*x*	*x*		*x*
ACpow_var	*x*	*x*		*x*

A 3 Summary of NCA for different operation modes
Chosen values are marked with a "x".

Feature	***Still Stand***	***Partial Load***	***Full Load***	***All Modes***
omega_min				
omega_max				
omega_mean				
omega_range				
omega_mode				
omega_std				
omega_var				
acc_x_min				
acc_x_max	x			
acc_x_mean	x	x	x	x
acc_x_range	x			
acc_x_mode				
acc_x_std	x	x	x	x
acc_x_var				
acc_y_min				
acc_y_max	x			x
acc_y_mean				
acc_y_range	x			x
acc_y_mode				
acc_y_std	x			
acc_y_var				
v_wind_min				
v_wind_max	x			
v_wind_mean		x		x
v_wind_range				
v_wind_mode				
v_wind_std		x		x
v_wind_var				x
v_dir_min				
v_dir_max				
v_dir_mean	x			x
v_dir_range				x
v_dir_mode	x			
v_dir_std	x	x		
v_dir_var				
omega_gen_min	x			
omega_gen_max	x	x		

Feature	Still Stand	Partial Load	Full Load	All Modes
omega_gen_mean				
omega_gen_range	x	x		x
omega_gen_mode				
omega_gen_std			x	
omega_gen_var				
air_density_min			x	
air_density_max				
air_density_mean				
air_density_range				
air_density_mode				
air_density_std				
air_density_var				
pitch_min			x	
pitch_max				
pitch_mean			x	
pitch_range		x		
pitch_mode				
pitch_std			x	
pitch_var				
ACpow_min		x		
ACpow_max				
ACpow_mean		x		x
ACpow_range	x	x		x
ACpow_mode				
ACpow_std	x		x	x
ACpow_var				

~~5~~6 Data availability

The high-frequency measurements from the SCADA and strain gauge sensors are not available due to confidentiality issue.

~~6~~7 Author contributions

The work was carried out by AM, based on his Master Thesis at the Wind Energy Technology Institute under the supervision of MS and TF. AM preprocessed the data for machine learning purposes, implemented feature selection techniques, modeled fatigue loads with neural networks and ran a sensitivity analysis. MS initiated the issue, ran the data gathering campaign and processed the raw data. All authors were involved in the development of the manuscript.

~~7~~8 Competing interests

The authors declare that they have no conflict of interest.

~~8~~9 Acknowledgements

We acknowledge financial support by Land Schleswig-Holstein within the funding program Open Access-Publikationsfonds.

~~9~~10 References

Boslaugh, S. and Watters, P. A.: Statistics in a nutshell, O'Reilly., 2008.

Cosack, N.: Fatigue Load Monitoring with Standard Wind Turbine Signals, University of Stuttgart. [online] Available from: https://d-nb.info/1009926721/34, 2010.

Cosack, N. and Kühn, M.: Ueberwachung von Belastungen an Windenergieanlagen durch Analyse von Standardsignalen.pdf, in AKIDA Tagungsband, 6. Aachener Kolloquium für Instandhaltung, Diagnose und Anlagenüberwachung, 14.-15.11.2006, pp. 277–283., 2006.

Cosack, N. and Kühn, M.: Prognose von Ermüdungslasten an Windenergieanlagen mittels Standardsignalen und neuronaler Netze.pdf, in DMK 2007 - Dresdner Maschinenelemente Kolloquium : 5. und 6. Dezember 2007, pp. 461–476., 2007.

DNV/Risø: Guidelines for Design of Wind Turbines - 2nd Edition., 2002.

Draper, N. R. and Smith, H.: Applied regression analysis, Wiley., 1998.

Goldberger, J., Roweis, S., Hinton, G. and Salakhutdinov, R.: Neighbourhood Components Analysis, 2005.

Han, J. and Kamber, M.: Data Mining: Concepts and Techniques., 2006.

Hazewinkel, M.: Encyclopaedia of Mathematics (set), 1st ed., Springer, Netherlands., 1994.

IRENA: Renewable Power Generation Costs in 2014, Bonn, Germany., 2015.

MathWorks: Fit Data with a Shallow Neural Network - MATLAB & Simulink - MathWorks, [online] Available from: https://de.mathworks.com/help/deeplearning/gs/fit-data-with-a-neural-network.html (Accessed 5 June 2019), 2019.

Melsheimer, M., Kamieth, R., Liebich, R. and Heilmann, Christoph, Grunwald, A.: No Title, in German Wind Energy Conference (DEWEK) Proceedings - 19th to 20th May 2015 in the Congress Centrum Bremen, pp. 1–4., 2015.

Noppe, N., Weijtjens, W. and Devriendt, C.: Modeling of quasi-static thrust load of wind turbines based on 1 s SCADA data, Wind Energy Sci., 3, 139–147, doi:10.5194/wes-3-139-2018, 2018.

Parsian, M.: Data algorithms : recipes for scaling up with Hadoop and Spark, “O’Reilly Media, Inc.”, 2015., 2015.

REN21: Renewables 2018 Global Status Report, Paris, France., 2018.

Schedat, M. and Faber, T.: Fatigue load reconstruction on wind turbine structures using structural health monitoring, in Proceedings DEWEK 2017, 13th German Wind Energy Conference, pp. 1–4., 2017.

Schedat, M., Faber, T. and Sivanesan, A.: Structural health monitoring concept to predict the remaining lifetime of the wind turbine structure, in Proceedings of the 24th Conference on the Domestic Use of Energy, DUE 2016, pp. 1–5, IEEE., 2016.

Seifert, J., Vera-Tudela, L. and Kühn, M.: Training requirements of a neural network used for fatigue load estimation of offshore wind turbines, Energy Procedia, 137, 315–322, doi:10.1016/j.egypro.2017.10.356, 2017.

Sharma, N. and Saroha, K.: Study of dimension reduction methodologies in data mining, in International Conference on Computing, Communication and Automation, ICCCA 2015, pp. 133–137, IEEE., 2015.

Smolka, U. and Cheng, P. W.: On the Design of Measurement Campaigns for Fatigue Life Monitoring of Offshore Wind Turbines, in The Twenty-third International Offshore and Polar Engineering Conference, vol. 9, pp. 408–413, International Society of Offshore and Polar Engineers, Anchorage, Alaska, USA,., 2013.

Vera-Tudela, L. and Kühn, M.: On the Selection of Input Variables for a Wind Turbine Load Monitoring System, Procedia Technol., 15, 726–736, doi:10.1016/j.protcy.2014.09.045, 2014.

Vidal, R., Ma, Y. and Sastry, S. S.: Principal Component Analysis, pp. 25–62., 2016.

Ziegler, L., Gonzalez, E., Rubert, T., Smolka, U. and Melero, J. J.: Lifetime extension of onshore wind turbines: A review covering Germany, Spain, Denmark, and the UK, Renew. Sustain. Energy Rev., 82, 1261–1271, doi:10.1016/j.rser.2017.09.100, 2018.