

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.
- 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.

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"
- P.1 abstract: consider including quantitative results (e.g. errors between measured and estimated DEL do not exceed x%)
- (*) P.4 line 9-10: Some more information on the turbine would be appreciated, in particular rated power.
- (*) 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
- P.5 line 9-11: the number used as n_{eq} is usually given too
- P.5 line 14-16: How are the outliers detected? What were the (normalized) limit values to detect them?
- P.5 line 19: Can you give some more explanation on the descriptive statistic "mode"?
- P.6 Figure 1: nice figure, does help to understand the methodology
- 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?
- 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?

- (*) 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?
- 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.
- 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?
- P.7, line 12: more explanation about p-value and F-statistic would be appreciated.
- P.7 line 22-23 "the output is then predicted by applying a function": What kind of function?
- P.7 line 27: how are the weights decided for this paper?
- P.7 line 28: observations = features ? This is a bit confusing, since "observations" might also be used for different measurements (in time)
- (*) 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.
- P.9 line 19: a figure showing results about the collinearity would be helpful for this discussion
- P.9 line 22-23 "Many of the remaining variables ... model": an example with specific resulting values would be helpful
- P.10 line 6: How does stepwise regression avoid multicollinearity?
- P.11 Section 3.1.2: Discussion about PCA seems to be missing
- (*) 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.
- 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
- 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.
- P.12 line 11-13: these datapoints (outliers) are not visible in the figure. Maybe consider a different visualization
- P.12 line 10-11 and p.13 Figure 4: Personally I don't see the added value of this figure and discussion.
- (*) 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?
- P.12 line 16: I cannot find the result 0.99486 in the figures or tables
- P.14 Figure 5: It's not clear to which data sets the figures exactly correspond. Add sublabels please.

- 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?
- 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.
- 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?
- 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?
- P.16 line 11: is the dataset increased consecutive in time or by randomly picking data from the entire dataset of one year?
- Section 3.3: An additional conclusion could be made: From the introduction I understood Smolka and Cheng did a similar analysis. How do your results compare to theirs?
- (*) 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.
- 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.
- (*) 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?

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
- 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.
- 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.
- 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)
- P.6 line 4: Considered to add also the percentage of remaining data after the removal of missing data.
- P.8 line 7: typo "squared"

- P.9 line 25: It might be helpful to clearly state that from this point the second technique, PCA, is discussed.
- P.10, Figure 2: add a line at 99% to increase visibility
- P.11 line 8: typo "datasets is ~~the~~ again the"
- P.13 line 5: sentence can be missed very easily
- P.14 line 10: reference to results is missing
- P.16 line 14: an equivalent number in time (weeks, months) is easier to interpret than 10241 observations