Estimation of Damage Equivalent Loads of Drivetrain of Wind Turbines using Machine Learning

O Kamel[1,2], S Hauptmann[1] and C L Bottasso[2]

[1] MesH Engineering GmbH, Stuttgart, Germany
[2] Chair of Wind Energy, Technical University of Munich, Germany

E-mail: omar.kamel@mesh-engineering.de

Abstract. In the recent years machine learning techniques have attracted the attention of wind energy community to make use of the large amount of available data produced from the running wind turbines. These modern wind turbines are typically equipped with measurement systems and sensors that can provide a wealth of information about the operating conditions of the machine. Nevertheless, not all the acquired raw data can be used effectively to enhance the operation of a turbine. This work addresses the question of estimating the damage equivalent loads (DEL) of different components of a drivetrain. The estimation is based on low frequency sampled typically available SCADA measurements. Typical SCADA measurements that are used as input for the estimation model are generator rotational speed, low speed shaft torque and generator torque as well as, wind speed and direction. Several machine learning methods as random forests (RF), support vector machines (SVM), linear regression (LR), decision trees and neural networks (NN) were developed, exhibiting different behavior for each approach. The qualitative and quantitative performance of each algorithm are evaluated and compared against each other. Furthermore, analysis of importance of the input features is presented.

1. Introduction

Machine learning can play a very significant role in enhancing the asset performance in terms of, for example lifetime extension, reduction of maintenance costs, reduction of downtime, prediction of components failure. These goals can be achieved by exploiting the raw data available from the installed measurement and control systems on the assets. These measurements in themselves can not produce an added value to the asset status, unless they are extracted and processed in an intelligent manner in order to reveal hidden or not directly usable information. Insights in the behavior of the system can be discovered using machine learning and data analytics techniques. Moreover. data analytics algorithms can be incorporated in digital twins frameworks, which can be combined with the asset management system to achieve the desired goals.

SCADA measurements are diverse, i.e. there are mechanical measurements such as positions, velocities, accelerations. There are also thermal and acoustic measurements that are typically available from SHM and SCADA systems installed within the turbines. These measurements are typically abundant in quantity and source as they can be available starting from 1 second interval up to 10 minutes averaged data [1]. The feasibility of using SCADA measurements to reflect the structural status of the asset is investigated in [2]. The author also developed a framework for load estimation using standard SCADA measurements. The proposed approach exhibited acceptable performance using both data-based and physics-based methods. Several use cases were demonstrated by using different artificial intelligence techniques with SCADA measurements to enhance the performance of the assets under investigation [3]. The possibility of early failure and anomalies during operation utilizing SCADA measurements was investigated in [4]. The paper showed successful detection of early failures of gearboxes in wind turbines, and failure of generator bearings in direct drive wind turbines. Kalman filters and state estimators were used in several contributions to predict the load history of mechanical components based on structural health monitoring (SHM) systems [5, 6]. These state estimators can be based on simplified physics-based models or data-driven models.

The authors in [7] developed a statistical approach to reconstruct the torque histograms based on SCADA measurements of power output and rotor speed. The torque histograms are then incorporated in fatigue calculation algorithm. This approach yielded reduction in prediction error with 10% compared the state-of-industry algorithms. The approach proposed in [8] used a linear regression model to quantify the load distribution on blade bearings of wind turbines. The authors concluded that the approach showed discrepancies because the used methods are based on research on smaller bearings not taking into effect the realistic behavior of wind turbines blade bearings.

Neural networks were used to develop a data-based model to estimate the thrust loads on the rotor from SCADA measurements [9]. These neural networks were then incorporated into a damage quantification algorithm.

The work proposed in this publication offers a straight-forward and easy to implement approach fully based on low-frequency SCADA measurements overcoming the drawbacks of the physics-based and grey-box models found in literature.

2. Objectives

The increase in interest to extend the lifetime of mechanical components, which means reducing down-time due to faults or maintenance, requires finding unorthodox approaches to exploit the available data in order to reach the specified objectives of the owners and operators of assets. The aim of this contribution is to formulate and demonstrate data-driven models based on machine learning that should be able to estimate DELs endured by mechanical components of a drivetrain.

The investigated problem assumes the need of a DEL estimator in the drivetrain during operation. Physics-based approaches, such as finite element and/or multibody dynamical models, can not offer such possibility due to their high computational demand. Therefore there is a need to develop a data-driven approach using the readily available measurements from the installed sensors in the wind turbine. DELs resulting from the vertical reaction of the left torque arm are considered in this study.

3. Methodology

The concept assumes that the DEL can be expressed as a function of the standard SCADA measurements: generator speed, air-gap torque, torque at low-speed shaft, pitch angle, wind speed and direction. Additionally, SCADA systems typically provide statistical quantities as mean and standard deviation of measurements. The SCADA system provides the measurements in a discrete manner ($\Delta t|_{k-1\rightarrow k} = 5$s in this study). The formulation behind the data-driven estimation of DELs writes

$$
\begin{aligned}
DEL^{k-1\to k} = f(&T_{LSS}^{k}, T_{LSS}^{k-1}, \mu_{T_{LSS}}^{k-1\to k}, \sigma_{T_{LSS}}^{k-1\to k},\\
&T_{Gen}^{k}, T_{Gen}^{k-1}, \mu_{T_{Gen}}^{k-1\to k}, \sigma_{T_{Gen}}^{k-1\to k},\\
&\omega_{Gen}^{k}, \omega_{Gen}^{k-1}, \mu_{\omega_{Gen}}^{k-1\to k}, \sigma_{\omega_{Gen}}^{k-1\to k},\\
&\theta^{k}, \theta^{k-1}, \mu_{\theta}^{k-1\to t}, \sigma_{\theta}^{k-1\to t},\\
&v_{wind}^{k}, v_{wind}^{k-1}, \mu_{v_{wind}}^{k-1\to k}, \sigma_{v_{wind}}^{k-1\to k},\\
&\phi_{wind}^{k}, \phi_{wind}^{k-1}, \mu_{\phi_{wind}}^{k-1\to k}, \sigma_{\phi_{wind}}^{k-1\to k})
\end{aligned}
\tag{1}
$$

where T_{LSS} is the torque on low-speed shaft, T_{Gen} is the air gap torque applied on the generator shaft, ω_{Gen} is the generator rotational speed, θ is the pitch angle, v_{wind} is the wind velocity at hub height, ϕ_{wind} is the wind direction at hub height, $(\bullet)^{k-1}$ is the SCADA measurement at the beginning of the interval, $(\bullet)^{k}$ is the SCADA measurement at the end of the interval, $\mu_{\bullet}^{k-1\to k}$ is the mean value of the respective quantity in the interval $k-1 \to k$, $\sigma_{\bullet}^{k-1\to k}$ is the standard deviation of the respective quantity in the interval $k-1 \to k$, $DEL^{k-1\to k}$ is the damage equivalent load acting on the component in the interval $k-1 \to k$.

3.1. Approach and Workflow

The proposed approach is described in the flow chart of Fig. 1. It should be mentioned that the proposed approach shall be implemented within the asset management system of wind turbines where SCADA measurements would be directly fed to the implemented software. This work is based on simulations, and therefore SCADA measurements are mimicked by using high-fidelity simulation models of an exemplary wind turbine.

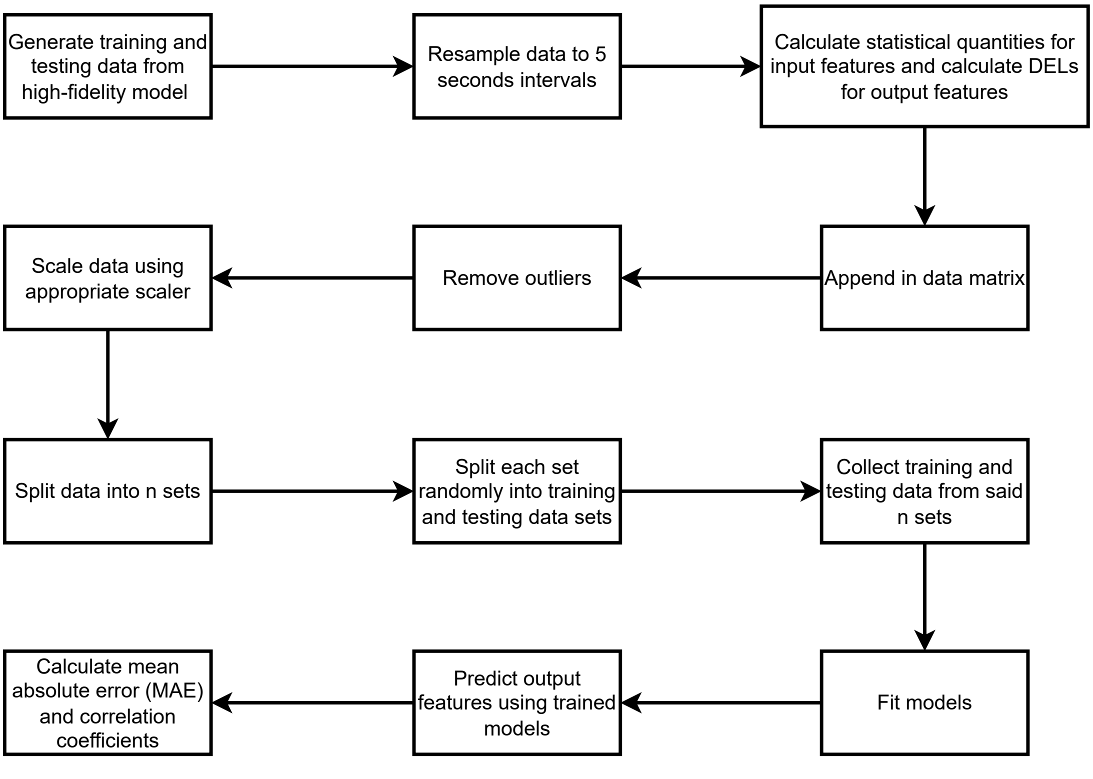
Generate training and testing data from high-fidelity model
Resample data to 5 seconds intervals
Calculate statistical quantities for input features and calculate DELs for output features
Append in data matrix
Remove outliers
Scale data using appropriate scaler
Split data into n sets
Split each set randomly into training and testing data sets
Collect training and testing data from said n sets
Fit models
Predict output features using trained models
Calculate mean absolute error (MAE) and correlation coefficients

Figure 1. Proposed Workflow for DELs Estimation

3.2. Generation of Training and Testing Data

The investigated wind turbine is the IEA 3.4 MW reference wind turbine [10]. The simulation model of the wind turbine is developed using the aeroservoelastic simulation software Simpack, where a high-fidelity multibody simulation (MBS) model is developed (cf. Fig. 2 and 3). The developed multibody simulation model in Simpack considers detailed modeling of flexible elements such as low-speed shaft and flexible teeth contact between gears [11]. Inflow aerodynamics is also considered using coupling of blade element momentum (BEM) solver `Aerodyn` [12].

Figure 2. Isometric View of Wind Turbine

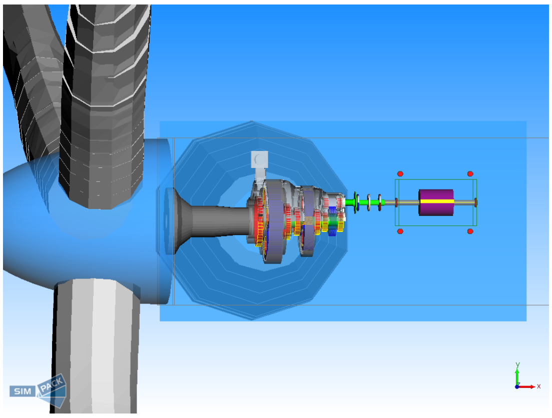

Figure 3. Top View of Wind Turbine showing Drivetrain

Several load cases were generated in Simpack according to IEC 61400-1 [13]. IEC standard defines different loadcases for evaluating the design of wind turbines. As the focus here is generating data that are relevant to calculating DELs on mechanical components, only power production loadcases were considered, namely design load cases (DLC) 1.2 and 1.3.

DLC 1.2 simulates a wind turbine operating under normal turbulence model (NTM), while DLC 1.3 simulates a wind turbine operating under extreme turbulence model (ETM). Each DLC realization produces time histories of the output of interest (vertical reaction on the left torque arm), through which DELs are calculated using rainflow counting algorithm [14]. Each DLC reveals different probabilistic distribution of the endured DELs on the components. The probabilistic distribution of DELs due to vertical reaction on the torque arm is demonstrated in figure 4, where the correlation between the distribution of DELs, wind speed and the turbulence models is evident.

The data provided by Simpack is sampled at high frequency of $f = 1000\texttt{Hz}$. SCADA systems typically provide measurements at much lower frequency. Therefore the DELs are evaluated in predefined window of $\Delta t = 5\texttt{s}$ to imitate SCADA system behavior.

Afterwards, the statistical quantities (cf. Eq. 1) are calculated for the input features: $T_{LSS}, T_{Gen}, \omega_{Gen}, v_{wind}, \phi_{wind}$ for each window between the discrete measurements. The output feature for the proposed approach is the DEL, which is an indirect quantity that has to be calculated firstly. The `Fatpack` package [15] is used for calculating the rainflow-matrix and perform the rainflow counting to calculate the damage equivalent loads. The calculation of DELs from the rainflow-matrices according to [14, 2, 16] writes

$$\Delta S_{eq,N_{ref},m} = \sqrt[m]{\frac{\sum_{i=1}^{n} \Delta S_i^m N_i}{N_{ref}}} \tag{2}$$

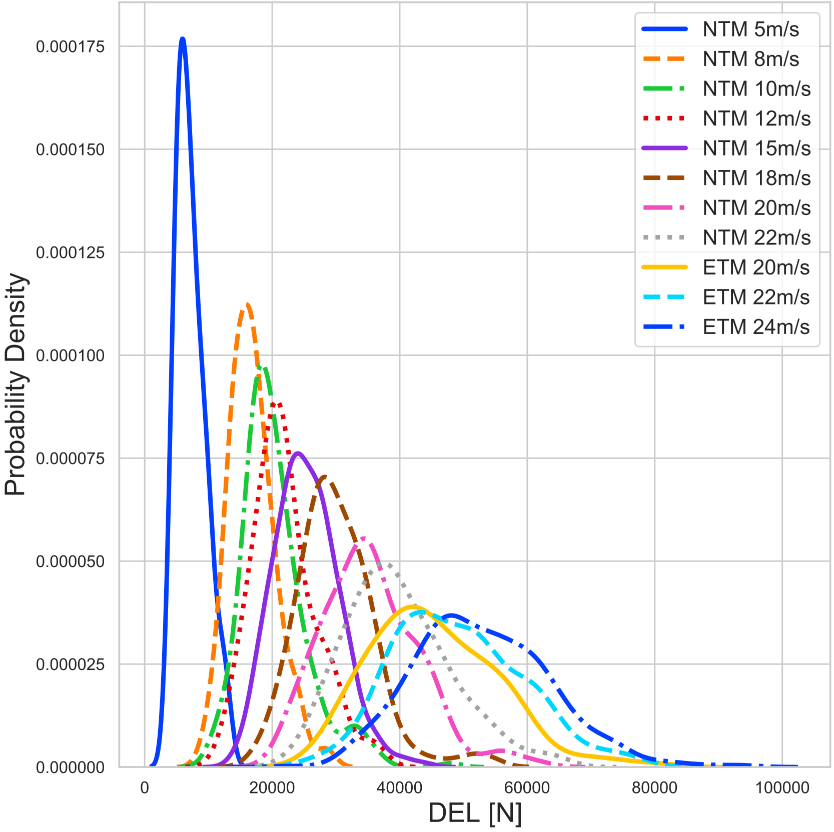

Figure 4. Probabilistic Distribution of DELs for DLC 1.2 and 1.3

where n is the number of load ranges, m is the material constant from Wöhler's curve, N_i is the number of rainflow cycles for the i^{th} stress range, N_{ref} is the reference number of rainflow cycles for which damage occurs at the stress range, ΔS_i is i^{th} stress range, $\Delta S_{eq,N_{ref},m}$ is the accumulated linearly damage equivalent load for the considered load timeseries.

3.3. Machine Learning Workflow

After generation and preprocessing of the data from the simulation, these data should be prepared with a different perspective in order to be used with the machine learning workflow. It is clear from figure 4 that the distribution of DELs is not following a normal distribution, therefore caution should be taken when dealing with such skewed data, especially in the extreme region of distribution. Outliers are removed from the data set to be able to fit the learning algorithms in a better way, as the learning algorithms commonly perform inadequately in the extreme region of the given data set. Interquartile Range Method (IQR) is used to remove the outliers within the data set (Fig. 5).

The data set is then scaled using `MinMax()` scaler in order to normalize the input and output features of the model. The scaled data set is then split into n random data subsets and each subset is split into training and testing randomly and then all the training and testing subsets are collected together in order to increase the randomness in the data set to overcome model overfitting.

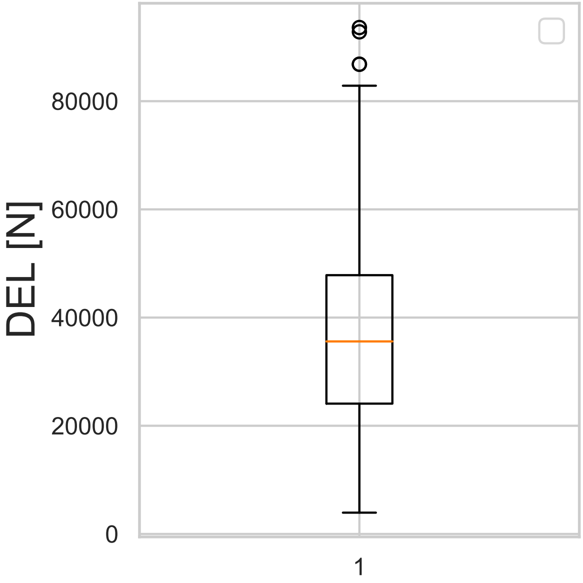

Figure 5. Box Plot of DELs Distribution of DELs for DLC 1.2 and 1.3

3.4. Investigated Algorithms

A library of regression algorithms were implemented and investigated in this work. Support Vector Regression (SVR), Gaussian Process Regression (GPR) with Dot Product and White Noise kernel, Decision Trees (both normal and ensemble trees), Random Forest (RF) with 100 estimators, Linear Regression and finally deep Neural Network (NN) were developed. The extra trees are considered to be ensemble-averaged normal decision trees improving prediction accuracy and protecting the model from overfitting. Validation loss of data-driven models is defined as the mismatch between ground truth data points and predicted data points evaluated on the validatid data set during model training. The convergence curve of the validation loss during training of neural networks is exemplarily in Fig. 6 demonstrated.

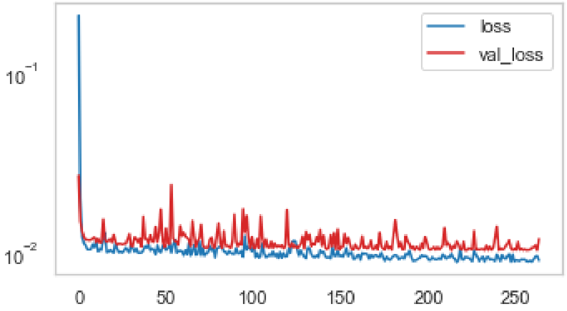

Figure 6. Validation Loss of Neural Network (exemplary)

4. Results

The evaluation of the proposed approach is realized by examining several evaluation criteria. The criteria assessed in this contribution are the correlation coefficients and mean absolute error of predictions against the reference ground truth.

Table 1. Mean Absolute Error of the Investigated Algorithms

SVR	GPR	Tree	Extra Tree	RF	LR	NN
0.0741	0.0743	0.0943	0.1028	0.0731	0.0743	0.0801

Table 2. Pearson's Correlation Coefficients

SVR	GPR	Tree	Extra Tree	RF	LR	NN
88.8%	88.8%	81.8%	78.3%	89%	88.8%	89.1%

4.1. Correlation

Figure 7 illustrates the correlation between predictions of the investigated methods and the ground truth. The predictions deviate noticeable towards the extreme region because there is much less number of observations in the extreme range (cf. Fig. 5 and 4). The mean absolute error of the predictions lies below 10% for all methods except for extra randomized decision trees (Table 1). Pearson's correlation coefficients were evaluated for all algorithms in Table 2. The coefficients are greater than 80% except for extra randomized trees. Fig. 8 shows the probabilistic distribution of the predicted DELs against the distribution of the ground truth data. Neural networks, SVR and LR exhibited over-estimation around the regions of $DEL = 0.2\texttt{N/N}$ and $DEL = 0.6\texttt{N/N}$ due to the sensitivity and overfitting of the algorithms with respect to the fed training data. Random trees and extra randomized trees yield very similar probabilistic distribution to the training data.

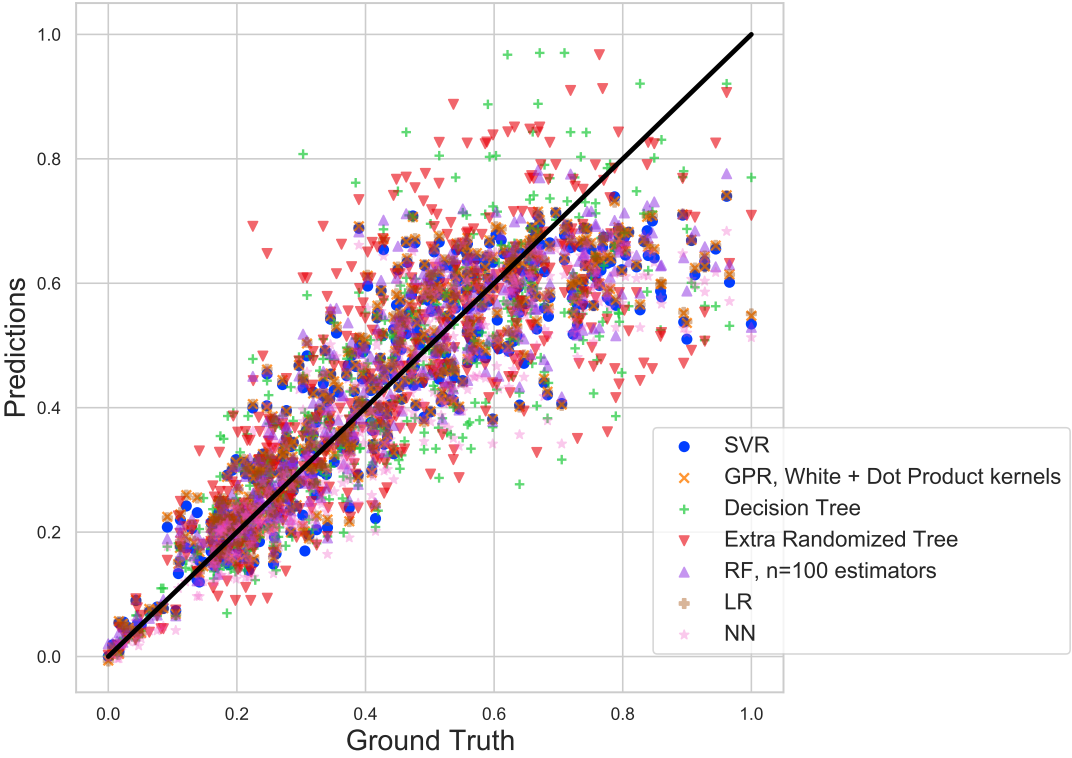

Figure 7. Correlation between Predictions and Ground Truth for the Test Data Set

The uncertainties associated with machine learning approaches arise due to the quantity of the training data and the random nature of the algorithms. There is a limitation on generating

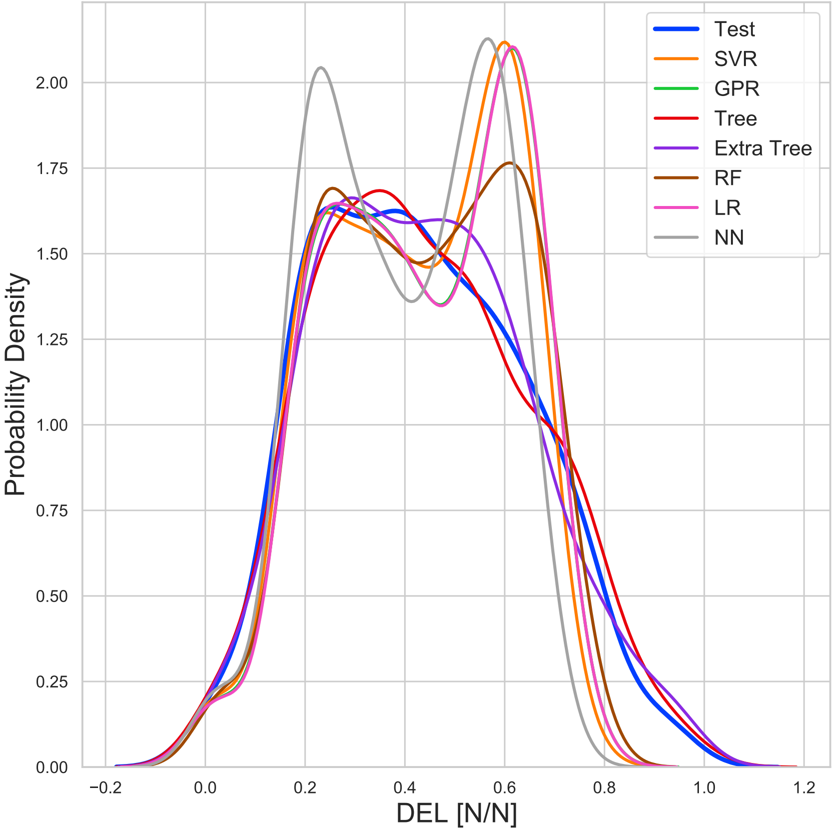

Figure 8. Probabilistic Distribution of Predicted DELs of each of the investigated algorithms

the training data due to the high computational cost of the simulation model (16 computational hours for 660 seconds of simulation). The randomness arises from the batching of training and testing data and the heuristic nature associated with the fitting of the hyperparameters of machine learning models. These uncertainties lead to slightly different performance of the models after each run of the training process. Nevertheless, the differences in models performance are not of quality-affecting nature, as the predicted DELs followed similar statistical distribution after each run of training.

4.2. Features Importance

One of the key measures to assess the structure of regression models is to investigate the importance of the input features in order to identify the importance of each input feature and how to retune the model disregarding less important features to increase model computational efficiency and prediction performance. Fig. 9 demonstrates the relative feature importance of one of the regression methods used (Random Forests). It is evident that not all features have equal importance in contributing to the prediction of the model.

5. Conclusion

This paper proposed an approach to develop a data-driven methodology to estimate damage equivalent loads of mechanical components of drivetrains of wind turbine using standard SCADA measurements. The data-driven method was trained and tested using artificial SCADA measurements generated using aeroservoelastic simulation of a reference wind turbine.

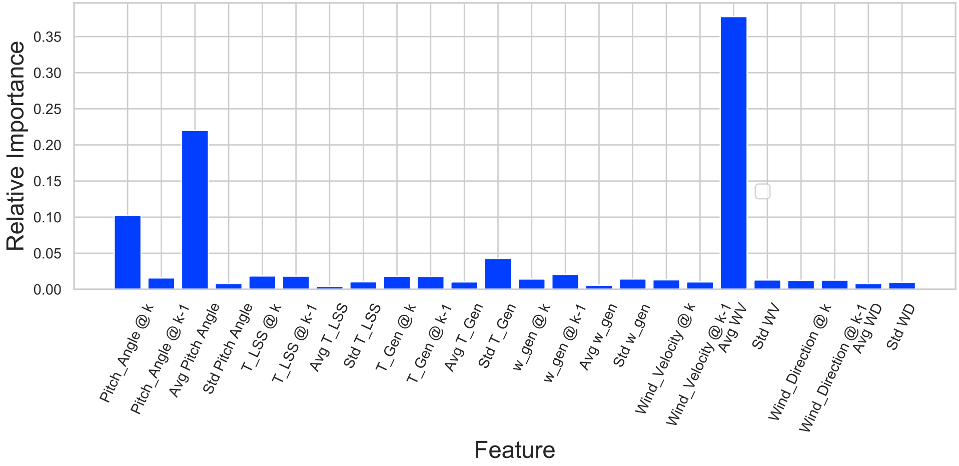

Figure 9. Feature Importance for Random Forest Model

The proposed workflow resulted in acceptable correlation using the given set of SCADA measurements.

Several enhancements can be complemented to the proposed approach, namely increasing the size of training data sets by including more load cases especially the production loadcases. Increasing the size of training data set can help generalize the trained models to be able to capture more subregions of the data spectrum.

References

[1] Cevasco D, Tautz-Weinert J, Kolios A J and Smolka U 2020 *J. Phys.: Conf. Ser.* **1618** 022063 ISSN 1742-6596
[2] Nicolai Cosack 2010 *Fatigue load monitoring with standard wind turbine signals* PhD Thesis University of Stuttgart
[3] Dimitrov N and Natarajan A 2019 *J. Phys.: Conf. Ser.* **1222** 012032 ISSN 1742-6596 URL `https://iopscience.iop.org/article/10.1088/1742-6596/1222/1/012032`
[4] Letzgus S 2015 *SCADA-Data Analysis for Condition Monitoring of Wind Turbines* M.Sc. University of Stuttgart Stuttgart
[5] Branlard E, Jonkman J, Dana S and Doubrawa P 2020 *J. Phys.: Conf. Ser.* **1618** 022030 ISSN 1742-6596
[6] Namura N, Muto K, Ueki Y, Ueta R and Takeda N 2021 *AIAA Journal* 1–11 ISSN 0001-1452
[7] Alvarez E J and Ribaric A P 2018 *Renewable Energy* **115** 391–399 ISSN 09601481
[8] Menck O, Stammler M and Schleich F 2020 *Wind Energ. Sci.* **5** 1743–1754 ISSN 2366-7451
[9] Santos F d N, Noppe N, Weijtjens W and Devriendt C 2020 *J. Phys.: Conf. Ser.* **1618** 022020 ISSN 1742-6596
[10] Pietro Bortolotti, Helena Canet Tarres, Katherine Dykes, Karl Merz, Latha Sethuraman, David Verelst and Frederik Zahle IEA Wind Task 37 on Systems Engineering in Wind Energy WP2.1 Reference Wind Turbines
[11] Dassault Systems Simpack Assistant 2020x
[12] P J Moriarty and A C Hansen 2005 AeroDyn Theory Manual URL `https://www.nrel.gov/wind/nwtc/assets/pdfs/ad-theory.pdf`
[13] IEC 2005 Wind turbines – Part 1: Design requirements
[14] ASTM 2017 Practices for Cycle Counting in Fatigue Analysis
[15] Gunnstein Thomas Frøseth fatpack URL `https://github.com/Gunnstein/fatpack`
[16] Thakur C 2021 *Estimation of Fatigue Damage in Wind Turbine Towers using Strains and SCADA measurements* M.Sc. Thesis Technical University of Munich