Prognosis of Wind Turbine Gearbox Bearing Failures using SCADA and Modeled Data

Arch Desai[1,2], Yi Guo[2], Shawn Sheng[2], Caleb Phillips[2], and Lindy Williams[2]

[1]*Texas A&M University, College Station, Texas, 77840, USA*
Archdesai.ad@gmail.com

[2]*National Renewable Energy Laboratory, Golden, Colorado, 80401, USA*
Arch.Desai@nrel.gov
Yi.Guo@nrel.gov
Shawn.Sheng@nrel.gov
Caleb.Phillips@nrel.gov
Lindy.Williams@nrel.gov

ABSTRACT

Predictive maintenance and condition monitoring systems for wind turbines have seen increased adoption to minimize downtime, reducing operation and maintenance costs. On today's wind power plants, the integrated supervisory control and data acquisition (SCADA) system provides low-frequency operational data that can be leveraged to quantify a wind turbine's health. The aim of this study is to utilize machine-learning techniques to predict axial cracking failures in wind turbine gearbox bearings up to 1 month ahead of time. The failures are assumed to have occurred when the investigated bearing was replaced. While current SCADA systems show the overall condition of a wind turbine, often they do not allow for the investigation of specific gearbox bearings' health. To enrich bearing fault signatures, additional data are computed through physics-based models using gearbox design information. Based on SCADA data, modeled data, and bearing failure log data from an actual wind plant, the performances of different machine-learning models on unseen data are then evaluated using industry-standard metrics, such as precision, recall, F1 score, and area under receiver operating characteristic curve (AUC). Results show the overall system performance enhancement in predicting bearing failure when modeled data are included with SCADA data. The reduction in terms of false alarms is about 50%, and improvement in terms of precision, F1 score, and AUC is about 33%, 12%, and 6%, respectively, based on the best performing modeling case in this study.

1. INTRODUCTION

Wind energy has been advancing rapidly as it plays a very significant part in clean-energy-based electric power generation (Kusiak & Li, 2011). However, the operation and maintenance (O&M) cost of a wind farm accounts for up to 30% of total energy cost (Fischer, Besnard, & Bertling, 2012), which can be reduced through continuous monitoring and successfully detecting incipient wind turbine failures. For this reason, predictive maintenance and condition monitoring systems are being implemented for O&M decision-making in the wind industry (Lau, Ma, & Pecht, 2012; Feng, Qiu, Crabtree, Long, & Tavner, 2013; Qiao, & Lu, 2015). These systems utilize a combination of statistics, data mining, and machine-learning-based techniques for fault diagnostics and prognostics, to assess wind turbine performance abnormalities and predict time to failure. The performance of these systems is improving with advances in data acquisition and signal processing technologies (Qiao, Zhang, & Chow, 2015; Zaher, McArthur, Infield, & Patel, 2009; Colone, Reder, Tautz-Weinert, Melero, Natarajan, & Watson, 2017).

Wind turbines operate in adverse weather conditions and their drivetrains undergo severe variable loading because of emergency shutdowns, varying wind speed, and fluctuations in energy demand (Yang, Tavner, Crabtree, Feng, & Qiu, 2013). One of the most important parts of a geared wind turbine drivetrain system is a gearbox that comprises various bearings, gears, and shafts. Most of the gearbox failures are related to the shaft bearings and result in very costly repairs and high downtime (Saidi, Ben Ali, Bechhoefer, & Benbouzid, 2017). Research has shown that cracks can develop in the gearbox bearings only within 3 years of a wind turbine's operation (Stadler & Stubenrauch, 2013). The failed bearings can damage surrounding components of the

gearbox, resulting in costly replacement of various components (Musial, Butterfield, & McNiff, 2007; Yang, Tavner, Crabtree, & Wilkinson, 2010). For these reasons, predicting wind turbine gearbox bearing failure is crucial.

A wide range of approaches for condition monitoring and fault prediction have been developed (Leite, Araújo, & Rosas, 2018). Kusiak and Verma (2012) analyzed the bearing faults considering over temperature events and using neural networks. Zhang (2018) put forward an automatic fault prediction method to predict main bearing fault using neural networks. The author predicts the bearing temperature using other features, such as active power output, ambient temperature, and turbine speed, and the fault is identified based on the prediction error. Koukoura et al. (2018) developed an approach to predict wind turbine planet bearing fault before 12 months, 6 months, and 1 month using a linear regression model. These approaches generally use historical data of wind turbines collected by a supervisory control and data acquisition (SCADA) system to identify patterns that lead to failure. The SCADA system typically records averages of sensor channels over 10-minute intervals to reduce storage and bandwidth. The data include various measurements, such as rotor speed, power, bearing temperature, and lubricant temperature (Zaher, McArthur, Infield, & Patel, 2009). These SCADA data show the overall condition of a wind turbine and can be leveraged to detect when the turbine's performance is degrading and to identify if a fault is developing. However, it becomes challenging to predict the failure of a specific wind turbine gearbox bearing, because the SCADA data are often not directly linked to any gearbox component. To bridge the gap, different features are calculated from SCADA data using physics-based models and the gearbox design. These modeled data, along with the SCADA data, are utilized for the bearing failure prediction, with an aim to improve the performance of current prognostics techniques.

The rest of the paper is organized as follows: Section 2 describes characteristics of data, the method to compute bearing-specific modeling data, and the data preprocessing technique we use. Section 3 presents strategies to address the class imbalance problem and machine-learning algorithms used for this study. In Section 4, we compare the performances of machine-learning models on historical bearing failure data using standard evaluation metrics. Finally, Section 5 summarizes our contribution to the field of wind turbine prognostics and discusses the areas of future work.

2. DATA DESCRIPTION AND PREPROCESSING

The data used in this study have been collected by a project partner from 12/01/2008 to 10/31/2018 at a wind farm located in Texas. The investigated data set contains a total of 13 1.5-MW wind turbines that have an identical gearbox configuration. Each shaft of the gearbox has two bearings (A and B) mounted on different axial locations, and all 13 turbines had encountered either high-speed shaft (HSS) or intermediate-speed shaft (IMS) axial cracking failure on bearing A or bearing B. The bearings were replaced, and lubricants of some turbines were also upgraded after the failure. Because the effect of installing new bearings on old components is beyond the scope of this study, we considered data up until the first failure of any bearing of the gearbox. The bearing replacement dates are assumed as their failure dates in this study, but they could be different from when the actual failures occurred or become detectable through instrumentation.

The clear separation of train and test data in the beginning is important as it helps avoid accidently sharing information of test data during model development based on train data. To assess the model performance across a wind turbine's life, we randomly selected 10 turbines for model training and used the remaining three turbines for testing. It is worth noting that the three turbines used for testing had different bearing failures, but we will be treating them as the same in our study. The rationale is that they share common SCADA data channels that could be indicative of bearing failures and the modeled data related to them is normalized. Figure 1 shows the timeline of different bearing failures of all the turbines and whether the data are used for model training or testing purposes.

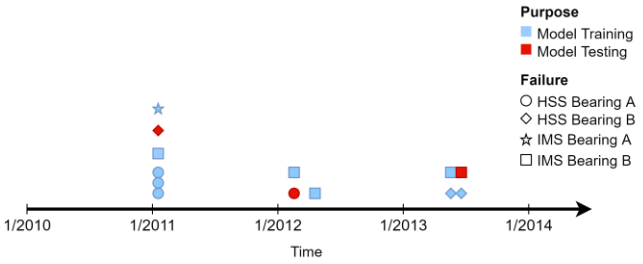

Figure 1. Timeline of bearing failures.

Table 1. SCADA channels.

SCADA Channel	Unit
Wind speed	m/s
Power	Watt
Rotor speed	m/s
Status code	-
Gearbox bearing temperature	Celsius
Gearbox oil temperature	Celsius
Nacelle temperature	Celsius
Ambient temperature	Celsius

The investigated SCADA data consist of 10-minute-interval averaged measurements of various sensors, as shown in Table 1. A total of 144 (6 per hour × 24 hours) rows of data are

recorded per day by a single turbine in the SCADA system. The gearbox bearing temperature is measured at another bearing (bearing C) on a high-speed shaft, which is different from the investigated bearings A and B and used to support the axial loads. The "status code" channel in the SCADA data shows the state of a wind turbine at a given time. During low wind speed, shutdowns, and maintenance activities, no power is produced by the wind turbine. Because these samples are out of the scope of our analysis, we filter them out and only consider the data when a turbine was in the running condition.

To represent the health condition of a bearing, additional data are calculated from the filtered SCADA data using physics-based models (Guo & Keller, 2020; Guo, Sheng, Phillips, Keller, Veers, & Williams, 2020). The various parameters for HSS bearings and IMS bearings in modeled data are shown in Table 2. An analytical roller sliding model (Guo & Keller, 2020) was used to calculate bearing roller and cage kinematic and sliding speed, which is important input for computing frictional energy generations on roller/raceway contact surfaces, as detailed in Guo et al., 2020. Bearing loads were calculated using a simple lumped-parameter model of high-speed and intermediate-speed shafts (Guo et al., 2020). Roller load distribution and deflections are estimated analytically using the approach in Harris & Kotzalas (2007). All these physics-based models require information on gearbox design and turbine operation history (SCADA data) as the input.

Table 2. Modeled data channels.

Modeled Data Channel	**Unit**
Bearing load	N
Bearing roller load	N
Roller deflection	mm
Rolling speed of roller and cage	rpm
Sliding speed of roller and cage	m/s
Slide-to-roll ratio	-
Frictional energy intensity	W/mm^2
Frictional energy	J

All time series data in Tables 1 and 2 (i.e., SCADA data and modeled data), are candidate features for this study. The Pearson correlation coefficient is calculated to determine linear correlation between each pair of features and to avoid multicollinearity. The features such as ambient temperature, bearing roller load, roller deflection, rolling speed, and sliding speed are dropped considering the collinearity threshold as 0.9. The selection of which colinear feature to drop is done by using domain knowledge. For example, nacelle temperature is kept as it is more related to the wind turbine than highly correlating ambient temperature. As a result of the high frequency of data collection, it is possible for SCADA data to contain noise and sensor errors. The outliers in the training data are detected using a 1.5-interquartile-range (IQR) method (Tukey, 1997), as it is considered robust against skewed data. Because the median values are not sensitive to outliers and show the central tendency in asymmetrical distributions, we replace detected outliers with median values. One thing to note is that the detected outliers are few and they occur randomly. Also, we do not see higher frequency of outliers when a turbine is about to fail.

Since we only filter the data when a turbine was producing power, we do not have a continuous stream of 10-minute frequency data and the total number of rows of data in any day can be between 0 and 144. With an assumption that the data do not change significantly in a single day, we aggregate the data generated on the same day and represent them with summary statistics, as shown in Table 3. Because the prediction horizon (1 month) is comparatively long, aggregated data help us understand meaningful trends and changes that will signify an impending failure. We use daily summary statistics of each feature for model training and testing.

Table 3. Summary statistics.

Summary Statistic	**Formula**
Minimum	-
Maximum	-
Length of data (N)	-
Mean ($\bar{x}$)	$\frac{\sum_{i=1}^{N} x_i}{N}$
Standard deviation (S)	$\sqrt{\frac{\sum_{i=1}^{N} \lvert x_i - \bar{x}\rvert^2}{N}}$
Root mean square (x_{rms})	$\sqrt{\frac{\sum_{i=1}^{N} {x_i}^2}{N}}$
Skewness	$\frac{\sum_{i=1}^{N} \lvert x_i - \bar{x}\rvert^3 \,/\, N}{S^3}$
Kurtosis	$\frac{\sum_{i=1}^{N} \lvert x_i - \bar{x}\rvert^4 \,/\, N}{S^4}$

Figure 2 shows the raw power values of a sample turbine from 2008-12-01 to 2008-12-07. Figure 3 shows how the same turbine's processed data (i.e., summary statistics) would look like for a same period.

The power a wind turbine produces at a given wind speed can be important to quantify its health. Figure 4 shows a typical power curve of a wind turbine (Sohoni, Gupta, & Nema, 2016).

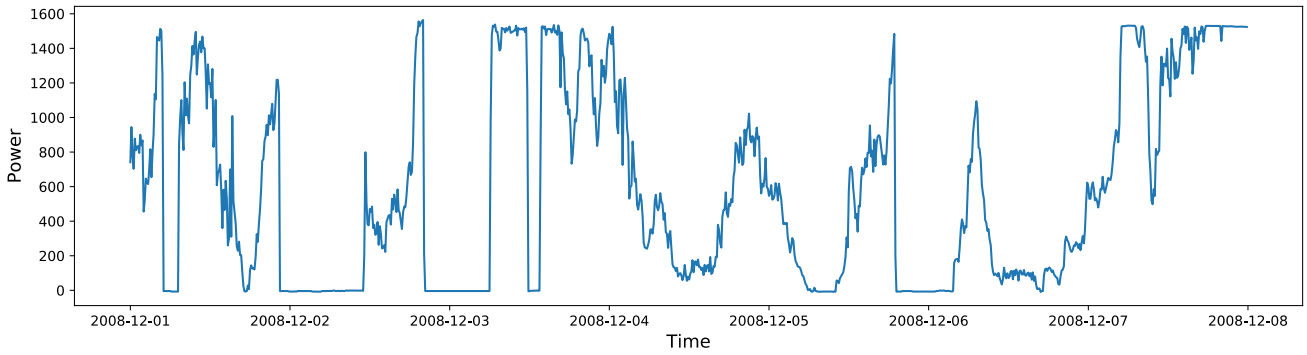

Figure 2. Raw power values of the sample turbine.

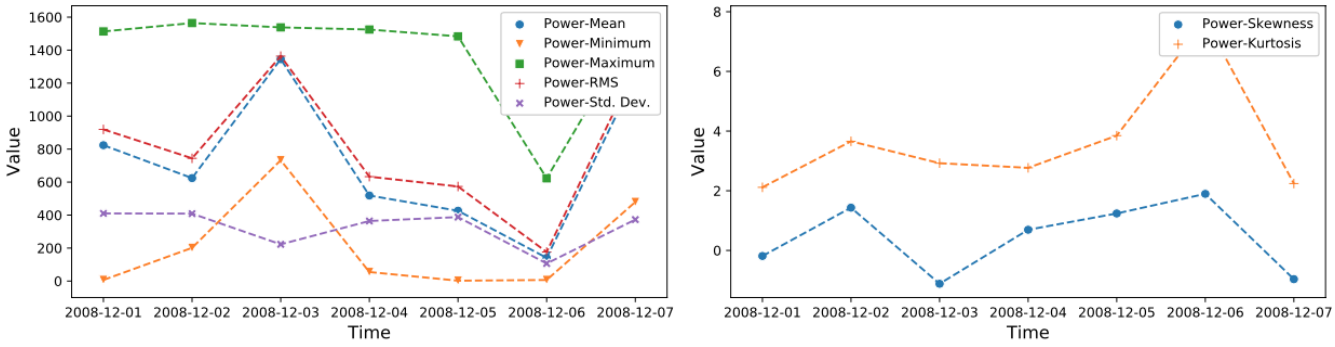

Figure 3. Summary statistics generated from raw power values of the sample turbine.

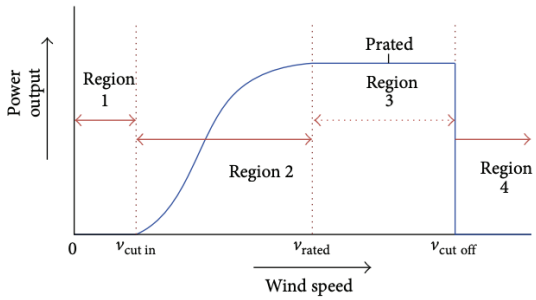

Figure 4. Wind turbine power curve.

A turbine operates at the rated power in Region 3, which is considered to be the best for its operation (Sohoni, Gupta, & Nema, 2016). In Region 3, load on the bearing is high and the bearing sliding is low, which results in less frictional energy. When a turbine is operating in Region 1 and early phase of Region 2, the bearing sliding is higher, thereby causing higher frictional energy (Guo & Keller, 2020). It has been proven that frictional energy causes white etching cracks in bench tests of bearing steel specimens (Gould & Greco, 2015). Hence, we calculate how many data points lie in each approximate region in a day using the wind speed limits given in Table 4. The regions are approximate because the upper wind speed limit (i.e., 6 m/s) for Region 1' is higher than a typical turbine's cut-in wind speed (e.g., 3 to 4 m/s), and the upper wind speed limit (i.e., 10 m/s) for Region 2' is lower than a typical turbine's rated wind speed (e.g., 12 to 15 m/s). They are defined based on whether the bearing is more prone to sliding and generates more frictional energy.

Table 4. Approximate power curve regions.

Region	**Wind Speed Limits**	
	Lower Limit	**Upper Limit**
Region 1'	0 m/s	6 m/s
Region 2'	6 m/s	10 m/s
Region 3'	10 m/s	25 m/s

Our aim is to predict bearing failure at least 1 month ahead of the actual failure (i.e., replacement date); therefore, the last 1 month of any wind turbine's data is labeled as "faulty" and earlier data are labeled as "healthy," as shown in Figure 5. The rationale behind such labeling is an assumption that the data from the last month before the failure contains a strong signal of bearing fault.

Figure 5. Data labeling.

3. METHODOLOGY

A supervised learning algorithm analyzes the training data and identifies a linear or nonlinear boundary separating healthy and faulty classes of the training data, which can then be used to predict a class of new instances. Therefore, we used supervised learning algorithms in this study. Because we mark only the last 1 month of data as "faulty," we have an extremely disproportionate ratio (~ 90:1) of observations in each class of the training data. Because of the class imbalance, algorithms become biased in favor of data with the majority class and ignore data with the minority class. Therefore, we utilize the following two techniques to address class imbalance and evaluate the performance of each in different models. We do not use undersampling techniques, as they lose a lot of information about the healthy turbine data.

- **Synthetic Minority Over-Sampling Technique (SMOTE):** SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) is an oversampling technique in which new examples from the minority class are synthetically generated. It selects two minority class instances randomly that are close in feature space and synthesizes a new sample at any point between the line joining them. This technique is better than random oversampling, which creates duplicate instances of a minority class that do not add any new information.
- **Cost-Sensitive Learning:** In cost-sensitive learning (Zadrozny, Langford, & Abe, 2003), we penalize the learning algorithm by increasing the cost of the classification mistake on the minority class. Because we have a class imbalance of about 90:1, we penalize the algorithm 90 times more if it wrongly classifies the faulty data as healthy.

To build bearing failure prediction models, we selected four algorithms: logistic regression (McCullagh & Nelder, 1989), random forest (Liaw & Wiener, 2002), extreme gradient boosting (XGBoost) (Chen & Carlos, 2016), and long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), which have the lowest to highest complexity, respectively.

Logistic regression is a linear classifier that finds a hyperplane in a feature space to separate observations according to their classes. The logistic regression transforms the output of a linear function using a logistic sigmoid function, such that it results in the probability value that can be mapped to a specific class. We utilize a logistic regression algorithm to find a linear boundary between healthy and faulty wind turbine data.

Random forest is an ensemble learning method that combines the outputs of a large number of decision trees and finds a nonlinear boundary in the feature space to separate classes. A decision tree is considered as a weak learner and often results in overfitting. In a random forest algorithm, decision trees are grown using only a random subset of features that reduces correlation among them. The final prediction of the observation is based on the voting of all decision trees, which outperforms any single decision tree and accounts for errors made by some.

XGBoost is a decision-tree-based ensemble algorithm. It uses a gradient-boosting framework in which decision trees are grown sequentially by accounting for errors made by prior ones. XGBoost also finds the nonlinear boundary in a feature space among classes. Further, it is better than other boosting algorithms, as it is optimized via parallel processing, tree-pruning, and regularization to avoid bias and overfitting.

LSTM is an artificial, recurrent, neural network architecture that is used to learn order dependence of sequential data. One LSTM unit contains input, output, and forget gates, which are used by a model to remember previous values. LSTM is considered to be the state-of-the-art algorithm for time series classification, as it efficiently accounts for the lagged values and is able to capture any underlying pattern in the sequential data. We utilize an LSTM network to predict wind turbine bearing failure at any given time using the past 1 month of sequential data.

SMOTE and most of the existing sampling techniques cannot consider the temporal structure of multivariate sequential data. Because LSTM handles imbalanced data and learns to classify accurately with correct cost function, we use only cost-sensitive learning to train LSTM models.

The gearbox bearing failure does not happen instantaneously and recent historical data are also very helpful for prognostics. Logistic regression, random forest, and XGBoost evaluate data points at a given time without bringing forward information from the past. Therefore, we add lagging variables of SCADA channels up to 30 days while training these models. These lag values are used to capture any underlying pattern in the last 1 month before the bearing failure. LSTM models inherently handle a sequence of past observations as an input; therefore, we do not add lag variables while training them.

All four algorithms are trained and tested using the following two sets of data. The list of final features is also shown.

1. **SCADA data**: Daily summary statistics of power, wind speed, rotor speed, bearing temperature, oil temperature, and nacelle temperature; daily number of data points lying in power curve Region 1', Region 2', and Region 3'.

2. **SCADA data and bearing-specific modeled data**: Daily summary statistics of power, wind speed, rotor speed, bearing temperature, oil temperature, nacelle temperature, bearing load, slide-to-roll ratio, frictional energy intensity, and frictional energy; daily number of data points lying in power curve Region 1', Region 2', and Region 3'.

To compare the performances of algorithms it is important to choose the right evaluation metric. For a typical binary classification model, a confusion matrix is normally used to evaluate its performance (shown in Figure 6).

		Prediction	
		Healthy	**Faulty**
Actual	**Healthy**	True Negatives (TN)	False Positives (FP)
	Faulty	False Negatives (FN)	True Positives (TP)

Figure 6. Confusion matrix.

The confusion matrix terms are defined as follows:

- **True Negatives**: Correct predictions of "Healthy" class samples
- **False Negatives**: "Faulty" class samples incorrectly predicted as "Healthy" (i.e., "missed alarms")
- **True Positives**: Correct predictions of "Faulty" class samples
- **False Positives**: "Healthy" class samples incorrectly predicted as "Faulty" (i.e., "false alarms").

Because we have a class imbalance, standard metrics such as accuracy and error rate are biased toward the data with the majority class. We rely on precision, recall, F1 score, and area under receiver operating characteristic curve (AUC) to evaluate all models in this study.

We can calculate precision, recall, and F1 score using the formulas shown in Table 5. We consider the turbine's last 1 month of data as positives and earlier data as negatives. Precision is used to minimize false alarms, whereas recall is used to minimize missed alarms. As the bearing's failure cost is high, we would prefer high recall with very few missed alarms. However, there is a trade-off between precision and recall, and an increase in recall comes at a cost of a decrease in precision. If there is less precision in a prediction model, we get many false alarms. Because gearbox bearings are not easily accessible, inspecting false alarms costs a lot as well. Therefore, we also use the F1 score to evaluate the overall performance of a prediction model, as it is a harmonic mean of precision and recall.

AUC is the area under receiver operating characteristic-ROC (Fawcett, 2004) curve. A ROC curve is a probabilistic curve that illustrates the trade-off between true positive rate (TPR) and false positive rate (FPR) at various classification thresholds. TPR and FPR can be calculated using formulas given in Table 5. The AUC provides a summary of the model's performance using a single number and it shows the model's ability to distinguish between positive and negative classes. The AUC can be between 0.5 and 1, and the higher the AUC, the better the model. When the AUC is 0.5, the model does not have class separation capacity. Since AUC is threshold-independent, it quantifies the model's performance holistically and we can utilize it to compare different models.

Table 5. Metrics.

Metric	**Formula**
Precision	TP/(TP + FP)
Recall or TPR	TP/(TP + FN)
F1 score	2 × Precision × Recall/(Precision + Recall)
FPR	FP/(FP + TN)

We adopt the best or recommended practices in the model development processes. To build logistic regression and random forest models, we use the Scikit-Learn library (INRIA, 2018), and use the XGBoost library (DMLC, 2019) for XGBoost models. Logistic regression, random forest, and XGBoost use a number of hyperparameters. We performed a randomized grid search to select the optimum hyperparameters that yield the best results using F1 score as a scoring metric. To build LSTM models, we used the keras library (Chollet et al., 2015), which is a deep learning framework. The LSTM models have two hidden layers, with 50 and 25 nodes in each, respectively. To determine a number of hidden nodes in each layer, we used a trial-and-error method. Because LSTM models are prone to overfit, we added two layers of dropout, with a 50% dropout rate to avoid overfitting. Based on the highest accuracy on validation data, we selected a final architecture.

4. RESULTS

Table 6 and Table 7 summarize the performances of models built using SMOTE and cost-sensitive learning as a class-balancing technique, respectively. We compare these two methods to address class imbalance in the data. The best method to remedy class imbalance is highly dependent on the data set (Weiss, McCarthy, and Zabar, 2007) and in this problem, cost-sensitive learning performs slightly better than oversampling using SMOTE for most of the algorithm and data combinations. As observed in Table 7, logistic regression models have the highest recall of 0.86 (SCADA data) and 0.85 (both SCADA and modeled data) with cost-sensitive learning, whereas LSTM networks have the highest precision of 0.52, F1 score of 0.57, and AUC of 0.97 using both SCADA and modeled data. Because the F1 score is a

harmonic mean of precision and recall, it is reduced significantly by poor precision in all models built using logistic regression, random forest, and XGBoost algorithms.

By adding 1 to 30 days of lag values while training logistic regression, random forest, and XGBoost models, feature space increases significantly, which results in overfitting observed by higher performance on training data and lower performance on test data. The reason for poor performance of logistic regression models might be that the data are not linearly separable. For tree-based models such as random forest and XGBoost, it has been observed that their performance is compromised in high-dimensional data (Jiang, Cui, Zhang, & Fu, 2018; Nguyen, Huang, & Nguyen, 2015). LSTM models are preferred for prognostics problems, as they can store the information about the recent historical data well and exploit the time dependency between them. Although LSTM models perform better, they do not offer model interpretability and feature importance. LSTM models also require a large amount of data and they are more computationally expensive.

Table 6. Performance summary of models built using SMOTE.

Model	**Algorithm**	**Data**	**Model Performance**			
			Precision	**Recall**	**F1 Score**	**AUC**
1	Logistic regression	SCADA	0.10	0.80	0.18	0.90
2		SCADA + modeled	0.11	0.79	0.19	0.91
3	Random forest	SCADA	0.08	0.70	0.14	0.79
4		SCADA + modeled	0.10	0.69	0.17	0.84
5	XGBoost	SCADA	0.10	0.73	0.18	0.89
6		SCADA + modeled	0.11	0.72	0.19	0.90

Table 7. Performance summary of models built using cost-sensitive learning.

Model	**Algorithm**	**Data**	**Model Performance**			
			Precision	**Recall**	**F1 Score**	**AUC**
1	Logistic regression	SCADA	0.11	0.86	0.22	0.88
2		SCADA + modeled	0.13	0.85	0.23	0.91
3	Random forest	SCADA	0.12	0.58	0.20	0.82
4		SCADA + modeled	0.14	0.57	0.22	0.85
5	XGBoost	SCADA	0.12	0.72	0.20	0.90
6		SCADA + modeled	0.13	0.71	0.22	0.91
7	LSTM	SCADA	0.39	0.75	0.51	0.92
8		SCADA + modeled	0.52	0.62	0.57	0.97

Our aim is to improve current gearbox bearing axial failure prognostic methods by adding modeled data that capture a bearing's fault signature to existing SCADA data. The results show that, for all four algorithms investigated with cost-sensitive learning, when modeled data are added, precision is improved, which increases the F1 score. The modeled data help reduce the number of false alarms by sacrificing relatively less recall. The overall system performance is also improved by an increase in AUC score.

We can visualize the performances of all models on unseen test data using the confusion matrix and ROC curves. Figure 7 and Figure 8 show the performance of LSTM models built using only SCADA data and modeled data along with SCADA data, respectively. When modeled data are used, we can see that performance is improved, as evidenced by the number of false alarms reduced from 105 to 52, about 50% improvement. By referring to the metrics in Table 7, the improvement in terms of precision is about 33%, and the sacrifice in terms of recall is about 17%, leading to an improvement in F1 score by 12% and in AUC by about 6%.

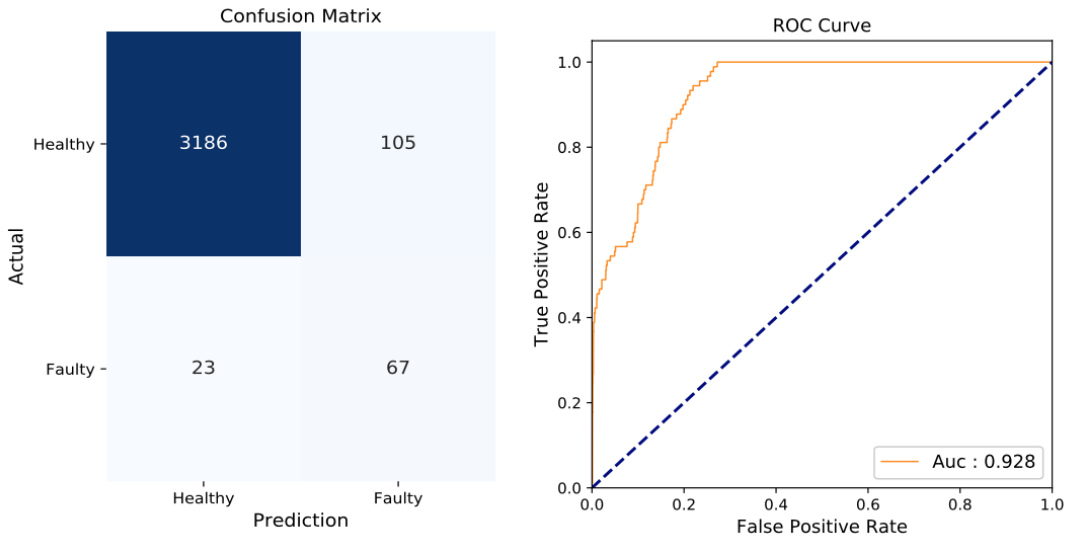

Figure 7. Performance of LSTM model built using SCADA data.

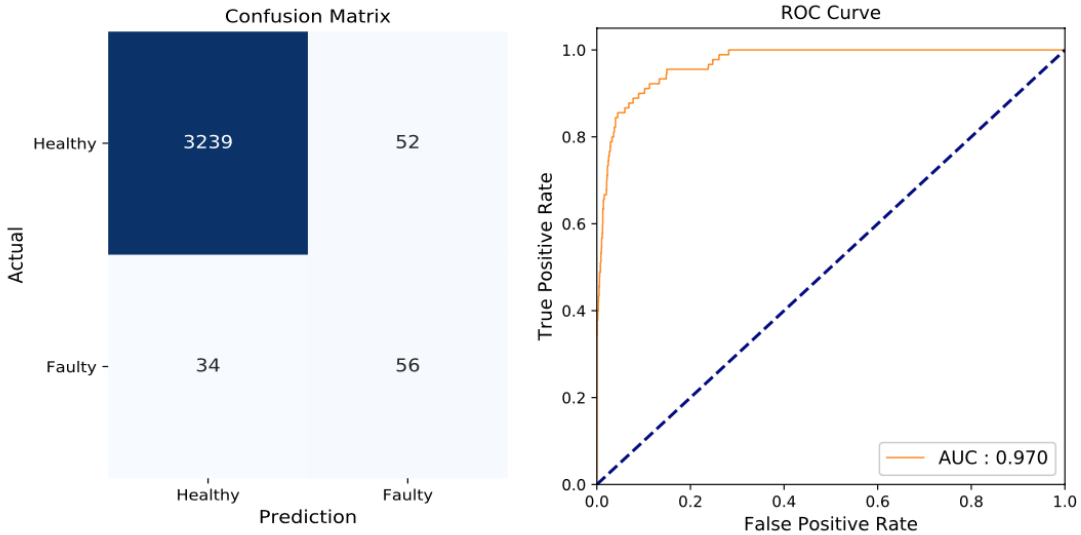

Figure 8. Performance of LSTM model built using SCADA and modeled data.

5. CONCLUSION

This study shows the potential of bearing-specific data computed using physics-based models and gearbox design to improve the existing bearing failure prognostics techniques that use only SCADA data. The modeled data help to reduce false alarms significantly and improve F1 score and overall AUC, about 50%, 12%, and 6%, respectively, in the best performing modeling case of this study. As there is a trade-off between precision and recall, the optimum model should be chosen by tuning the classification threshold on the ROC curve and considering the cost associated with missed and false alarms. In this study, we build a generalized model that can predict failure of any high-speed or intermediate-speed shaft bearing caused by axial cracking. However, overall performance can be improved if we make individual models for each bearing using SCADA and modeled data. We do not know when the bearing cracks start developing until they are visually detectable or through dedicated condition monitoring solutions. In this study, we try to detect signals in the last 1 month before a failure is assumed to have occurred when the bearing is replaced. Therefore, another area to further study this work would be to find the optimal time window to predict failure onsets of bearings that are deemed physically detectable and cost effective to support optimized maintenance decision-making.

ACKNOWLEDGEMENT

The authors acknowledge the project partner for sharing their wind power plant data and supporting our research.

This work was authored [in part] by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U. S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding was provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

REFERENCES

Chawla, V., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357. doi: 10.1613/jair.953

Chen, T., & Carlos, G. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi: 10.1145/2939672.2939785

Chollet, F., & others. (2015). Keras. Available at: https://keras.io

Colone, L., Reder, M., Tautz-Weinert, J., Melero, J.J., Natarajan, A., & Watson, S.J. (2017). Optimisation of data acquisition in wind turbines with data-driven conversion functions for sensor measurements. *Energy Procedia*, vol. 137, pp. 571-578. doi: 10.1016/j.egypro.2017.10.386

DMLC. (2019). Xgboost. Available at: https://xgboost.ai

Fawcett T. (2004). Roc graphs: Notes and practical considerations for researchers. *Machine learning*, vol. 31, pp. 1–38.

Feng, Y., Qiu, Y., Crabtree, C.J., Long, H. and Tavner, P.J. (2013). Monitoring wind turbine gearboxes. *Wind Energy*, vol. 16, pp. 728-740. doi:10.1002/we.1521

Fischer, K., Besnard, F., Bertling L. (2012). Reliability-centered maintenance for wind turbines based on statistical analysis and practical experience. *IEEE Transactions on Energy Conversation*, vol. 27 (1), pp. 184-195. doi: 10.1109/tec.2011.2176129

Gould, B., & Greco. A. (2015). The influence of sliding and contact severity on the generation of white etching cracks. *Tribology Letters*, vol. 60 (2), pp. 1-13. doi:10.1007/s11249-015-0602-6

Guo, Y., & Keller. J. (2020). Validation of combined analytical methods to predict slip in cylindrical roller bearings. *Tribology International*, vol. 148, p. 106347. doi: 10.1016/j.triboint.2020.106347

Guo, Y., Sheng, S., Phillips, C., Keller, J., Veers, P., & Williams, L. (2020). A methodology for reliability assessment and prognosis of bearing axial cracking in wind turbine gearboxes. *Renewable and Sustainable Energy Reviews, vol.* 127, p. 109888. doi: 10.1016/j.rser.2020.109888

Harris, T. A., Kotzalas, M. N. (2007). Essential concepts of bearing technology (Eds. 5). Boca Raton, FL: CRC Press

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, vol. 9 (8), p. 1735-80. doi: 10.1162/neco.1997.9.8.1735

INRIA. (2018). Scikit-learn. Available at: http://scikit-learn.org/stable.

Jiang, J., Cui, B., Zhang, C., & Fu, F. (2018). DimBoost: Boosting Gradient Boosting Decision Tree to Higher Dimensions. *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18). Association for Computing Machinery*, pp. 1363-1376. doi: 10.1145/3183713.3196892

Koukoura, S., Carroll, J., & McDonald, A. (2018). An insight into wind turbine planet bearing fault prediction using SCADA data. Proceedings of the European Conference of the PHM Society, vol. 4 (1)

Kusiak, A., Li, W. (2011). The prediction and diagnosis of wind turbine faults. *Renewable Energy*, vol. 36 (1), pp. 16-23. doi: 10.1016/j.renene.2010.05.014

Kusiak, A., Verma, A. (2012) Analyzing bearing faults in wind turbines: a data-mining approach. *Renewable Energy*, vol. 48, pp. 110-116

Lau, B., Ma, E., Pecht, M. (2012). Review of offshore wind turbine failures and fault prognostic methods. *Proceedings of the IEEE 2012 Prognostics and System Health Management Conference* (pp. 1-5), May 23-25, Beijing. doi: 10.1109/PHM.2012.6228954

Leite, G. d. N. P., Araújo, A. M., & Rosas, P. A. C. (2018). Prognostic techniques applied to maintenance of wind turbines: a concise and specific review. *Renewable and Sustainable Energy Reviews*, vol. 81 (2), pp. 1917-1925. doi: 10.1016/j.rser.2017.06.002

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), pp. 18-22.

McCullagh, P., & Nelder, J. (1989). Generalized linear models (Eds. 2). London: Chapman & Hall.

Musial, W., Butterfield, S., & McNiff, B. (2007). Improving Wind Turbine Gearbox Reliability. *Proceedings of the 2007 European Wind Energy Conference* (pp. 1–13), May 7-10, Milan, Italy.

Nguyen, T., Huang, J., Nguyen, T. (2015). Unbiased feature selection in learning random forests for high-dimensional data. *The Scientific World Journal*, vol. 2015, p. 471371. doi: 10.1155/2015/471371

Qiao, W., Lu, D. (2015). A Survey on Wind Turbine Condition Monitoring and Fault Diagnosis—Part I: Components and Subsystems. *IEEE Transactions on*

Industrial Electronics, vol. 62 (10), pp. 6536-6545. doi: 10.1109/TIE.2015.2422112

Qiao, W., Zhang, P., & Chow, M. Y. (2015). Condition monitoring, diagnosis, prognosis, and health management for wind energy conversion systems. *IEEE Transactions on Industrial Electronics*, vol. 62 (10), pp. 6533–6535. doi: 10.1109/tie.2015.2464785

Saidi, L., Ben Ali, J., Bechhoefer, E., & Benbouzid, M. (2017). Wind turbine high-speed shaft bearings health prognosis through a spectral Kurtosis-derived index and SVR. *Applied Acoustic*, vol. 120, pp.1–8. doi: 10.1016/j.apacoust.2017.01.005

Sohoni, V., Gupta, S., & Nema, R. (2016). A Critical Review on Wind Turbine Power Curve Modelling Techniques and Their Applications in Wind Based Energy Systems. *Journal of Energy, vol. 2016, p. 8519785. doi: 10.1155/2016/8519785*

Stadler, K., & Stubenrauch, A. (2013). Premature bearing failures in Industrial Gearboxes, SKF GmbH, Gunnar-Wester-Str, Schweinfurt, Germany, 97421.

Tukey, J. W. (1977). Exploratory data analysis. Pearson.

Weiss, G., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? Proceedings of the International Conference on Data Mining, pp. 35-41.

Yang, W., Tavner, P. J., Crabtree, C. J., Feng, Y., & Qiu, Y. (2013). Wind turbine condition monitoring: Technical and commercial challenges. *Wind Energy, vol.* 17, pp. 657–669. doi:10.1002/we.1508

Yang, W., Tavner, P., Crabtree, C., & Wilkinson, M. (2010). Cost-Effective Condition Monitoring for Wind Turbines. *IEEE Transactions on Industrial Electronics*, vol. 57 (1), pp. 263-271. doi: 10.1109/TIE.2009.2032202.

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *Third IEEE International Conference on Data Mining* (pp. 435-442), Nov 22-22, Melbourne, FL, USA. doi: 10.1109/ICDM.2003.1250950

Zaher, A., McArthur, S., Infield, D., & Patel, Y. (2009). Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy*, vol. 12(6), pp. 574–593. doi: 10.1002/we.319

Zhang, Z. (2018). Automatic Fault Prediction of Wind Turbine Main Bearing Based on SCADA Data and Artificial Neural Network. *Open Journal of Applied Sciences*, vol. 08, pp. 211-225. doi: 10.4236/ojapps.2018.86018.

BIOGRAPHIES

Arch Desai is a master's student in Industrial Engineering at Texas A&M University, College Station, Texas. He received B.Tech in Mechanical Engineering from National Institute of Technology (SVNIT), Surat, India in 2017. He has interned at National Renewable Energy Laboratory and has worked as an R&D Associate Engineer at Dassault Systèmes. He is primarily interested in Predictive Maintenance.

Dr. Yi Guo is a senior scientist at National Renewable Energy Laboratory. At NREL, Yi is dedicated to research and development in dynamic modeling, reliability analysis, vibro-acoustics, data analysis, and design optimization of land-based and offshore wind turbines. She has been focused on using analytical and computational modeling approaches to model and analyze geared systems and has published numerous impactful journal publications. Prior to NREL, she pursued her Ph.D. at Ohio State University with specialization in dynamics, vibration, and acoustics of wind turbine and helicopter drivetrains. Yi is a member of AGMA, ASME, and Women of Wind Energy. Yi also enjoys teaching students knowledge and hands-on experiences in engineering and renewable energy fields.

Dr. Shawn Sheng is a senior research engineer at National Renewable Energy Laboratory. He has B.S. and M.S. degrees both in electrical engineering and a Ph.D. in mechanical engineering. Currently, he leads condition monitoring, reliability database, and wind plant operation & maintenance research at NREL. Shawn has a broad range of experience: mechanical and electrical system modeling and analysis; data sensing and sensor placement; signal processing; machine defect classification and level evaluation; machine life prognosis; multi scale modeling; traditional and intelligent control. He is a member of STLE and ASME. His work has been published in various journals, conference proceedings, and book chapters.

Dr. Caleb Phillips is a data scientist with the Computational Science Center at NREL. He comes from a background in computer science systems, applied statistics, predictive modeling, machine learning, and optimization. His work supports projects across disciplines at NREL, including advanced and alternative-fuel vehicle technologies, transportation systems, materials science, photovoltaics, strategic energy analysis, wind power, and energy efficient high-performance computing. He maintains an adjunct appointment at the University of Colorado, Boulder where he has also done work in sustainability science (food systems), medical research (orthopedics and wilderness medicine), human computer interaction (power control), and artificial intelligence (assistance systems).

Lindy Williams is a data scientist with the Computational Science Center at NREL. With years of industry and laboratory experience combined with a Master's in Statistics, Lindy has gained problem solving and engineering skills as well as love for the entire statistical analysis process from data cleaning to final results. She also comes from a teaching background from years of tutoring and data science mentoring where she has enjoyed fostering interest in the data science field and helping others identify their unique skills and how those skills can be applied to particular problems.