Billah, Md Masum; Saberi, Alireza Nemat; Hemeida, Ahmed; Martin, Floran; Kudelina, Karolina; Asad, Bilal; Naseer, Muhammad U.; Mukherjee, Victor; Belahcen, Anouar

Generation of Unmeasured Loading Levels Data for Condition Monitoring of Induction Machine Using Machine Learning

Published in:
IEEE Transactions on Magnetics

DOI:
10.1109/TMAG.2023.3312267

Published: 01/03/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Billah, M. M., Saberi, A. N., Hemeida, A., Martin, F., Kudelina, K., Asad, B., Naseer, M. U., Mukherjee, V., & Belahcen, A. (2024). Generation of Unmeasured Loading Levels Data for Condition Monitoring of Induction Machine Using Machine Learning. *IEEE Transactions on Magnetics*, *60*(3), Article 8201104. https://doi.org/10.1109/TMAG.2023.3312267

Generation of Unmeasured Loading Levels Data for Condition Monitoring of Induction Machine Using Machine Learning

Md Masum Billah[1], Alireza Nemat Saberi[1,2], Ahmed Hemeida[1,3], Floran Martin [1], Karolina Kudelina [4], Bilal Asad [4], Muhammad U Naseer [4], Victor Mukherjee [5], and Anouar Belahcen [1], *Senior member, IEEE*
[1] Department of Electrical Engineering and Automation, Aalto University, Espoo 00076, Finland, md.billah@aalto.fi
[2]ABB Motion System Drives, Helsinki 00380, Finland
[3] Department of Electrical Engineering, Cairo University, 12613, Al-Jizah, Egypt
[4]Department of Electrical Power Engineering and Mechatronics, Tallinn University of Technology, Tallinn 19086, Estonia
[5]Technology Center, ABB Motors and Generators, Helsinki 00380, Finland

This article presents a novel data augmentation method that generates feature values for unmeasured loading levels based on limited measured and simulated loading level data. The incorporation of offline simulated data in the augmentation framework and the mapping of the error distribution over the loading levels greatly reduce the dependency on including a large number of loading levels in the curve fitting process. Furthermore, the proposed method shows high potential to minimize the deviation between measured and simulated data at the feature level. The method is applied to the induction machine to generate feature values at 25% and 50% loading levels for healthy, one, two, and three broken rotor bars conditions. An excellent agreement is observed between the augmented and actual feature values calculated from the measured data at 25% and 50% loading levels. The inclusion of this augmented data in the training phase aids in resolving the generalization issue and enhancing the average classification accuracy of the XGBoost algorithm by 9.4% and 4.4% at 25% and 50% loading levels, respectively.

Index Terms—**Broken rotor bar, condition monitoring, data augmentation, induction machine, machine learning.**

I. Introduction

CONDITION MONITORING of induction machines (IMs) using machine learning (ML) models has gained significant popularity in recent decades [1], [2]. However, most supervised ML models require a large amount of labeled data to achieve high prediction accuracy [3]. The single-phase stator current signal is commonly utilized for feature extraction and ML model training due to its availability and low cost [4] [5]. Nevertheless, training an ML model solely with single-phase stator current data for a limited range of loading conditions can lead to generalization problems. In other words, it may perform poorly on test data for a specific loading level not included in the training dataset [6]. Collecting data for all loading conditions from the measurement setup is impractical due to operational constraints. To overcome this limitation, simulation data for various loading levels can be employed during the ML training phase [7], [8]. However, simulation data may not accurately replicate measured data due to uncertainties and noise. Therefore, we propose a new technique that augments feature values at intermediate loading levels by using four sets of measured and simulated loading levels data for healthy and broken rotor bars (BRBs) of IMs.

In [4], [5], a curve fitting technique is presented to interpolate unmeasured loading levels from the discrete wavelet transform (DWT) and matching pursuit (MP) based decomposition of single phase stator current signal. Our approach differs from [4], [5] in several ways. Firstly, the interpolation is performed directly on features computed from six loading levels of measured data in [4] [5]. However, the quality of the generated data depends on the interpolation accuracy, which heavily relies on the number of loading levels used in the fitting process. This poses a challenge when dealing with a limited number of loading levels, as the trend of the features over the loading levels may not be known in advance. The novelty of our approach lies in leveraging offline simulated data and interpolating the error between measured and simulated features values. This reduces the dependency on adding more loading levels in the fitting process, as the error can be interpolated using lower degrees of polynomials. Secondly, we combine the interpolation technique with the probability distribution function (PDF) to account for the variation of error between measured and simulated features values and the variation from one window signal to another. Thirdly, our approach avoids complex decomposition methods such as DWT and MP for the stator current signal. Instead, we compute time-domain features from the raw data and obtain frequency-domain features using the fast Fourier transform (FFT). Finally, to augment features values for intermediate loading levels, we construct a PDF using the corrected mean and standard deviation and draw samples from it.

Manuscript received December 1, 2012; revised August 26, 2015. Corresponding author: M. Shell (email: http://www.michaelshell.org/contact.html).

II. Methodology

A. Measured Data Acquisition

Figure 1 presents the measurement setup. A 4-pole, 400 V, 50 Hz, 7.5 kW, and Δ-connected induction machine is used as the test machine. The test machine features a stator outer diameter of 220 mm and rotor inner diameter of 45 mm. Artificial BRB faults are introduced by drilling one hole in each rotor bar. Another induction motor with a similar rating is controlled through an ABB industrial drive ACS600 to

provide a constant load torque. The tested induction machine is supplied with a voltage from the grid. The experiment is conducted for four classes: healthy, one BRB, two BRBs, and three BRBs. Each class consists of five loading levels: 0%, 25%, 50%, 75%, and 100%. The experiment is carried out for 60 seconds at each loading level with a sampling frequency of 20 kHz.

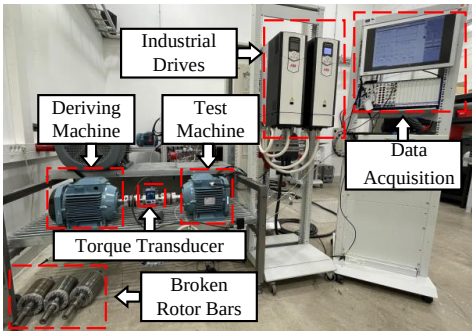

Fig. 1. Measurement setup.

B. Simulated Data Acquisition

A computationally efficient and precise magnetic equivalent circuit (MEC) model is used to simulate four classes: healthy, one, two, and three BRBs. The simulations are performed at five different loading levels: 0%, 25%, 50%, 75%, and 100%, while maintaining the same sampling frequency as the measured data. An example of the MEC model applied to BRB faults can be found in [9]. In our mesh-based MEC model, the stator and rotor nodes are coupled using Lagrange multipliers, and the non-linearity of the system equations is solved using the Newton-Raphson method. Furthermore, a time-harmonic MEC model is developed and integrated with a time-stepping MEC model to minimize transient effects. To simulate different loading conditions, the load torque is adjusted, and the speed is varied based on the equation of rotor motion. The BRB faults are incorporated into the rotor circuit equations by increasing the rotor resistance to an extremely high value, and only adjacent BRB faults are considered.

C. Proposed Data Augmentation Framework

An overlapping sliding window is employed with a window size of 16,000 data points from a single phase stator current. The window is shifted by 400 data points, and the total number of windows are 40 per loading level. For simplicity, five time-domain (TD) statistical features: TD_1 = mean, TD_2 = median, TD_3 = standard deviation, TD_4 = skewness, and TD_5 = kurtosis and five frequency-domain (FD) statistical features: FD_1 = mean, FD_2 = median, FD_3 = standard deviation, FD_4 = skewness, and FD_5 = kurtosis are considered. The time-domain features are computed directly from the raw windowed signal, while the frequency-domain features are obtained through the FFT. Each window yields one value, resulting in 40 values per loading level for each feature.

Figure 2 illustrates the workflow of the proposed data augmentation technique to generate the statistical features values for unmeasured loading levels. Initially, the error between the measured and simulated data for each feature is computed using four loading levels. Subsequently, a normal distribution is fitted to estimate the mean ϵ_μ and standard deviation ϵ_σ of the error for each feature corresponding to their loading level. Polynomial functions, namely linear, quadratic, and cubic are used to interpolate ϵ_μ and ϵ_σ versus their corresponding loading levels. These polynomial functions are utilized to estimate the mean ϵ^*_μ and standard deviation ϵ^*_σ of the error for unmeasured loading levels. The estimated mean ϵ^*_μ and standard deviation ϵ^*_σ of the unmeasured loading level is added to the mean $F_{s\mu}$ and standard deviation $F_{s\sigma}$ of features that are calculated from the simulated data corresponding to their unmeasured loading level. Furthermore, a normal distribution is fitted using the corrected mean and standard deviation of each feature, and samples are drawn for each feature of the unmeasured loading level. For a better illustration purpose, a hypothetical representation of a single feature value augmentation at 25% loading level is demonstrated in Fig. 3. The augmented feature values are validated by comparing them with the features computed from the measured data for the corresponding loading level.

D. Selected Supervised Machine Learning Algorithms

To perform the classification tasks, we have adopted six supervised ML algorithms: K-nearest neighbors (KNN), support vector machine (SVM) with radial basis function (RBF) kernel, decision tree (DT), random forest (RF), adaptive boosting (AdaBoost), and extreme gradient boosting (XGBoost) from scikit-learn ML library. The inputs of these ML models consist of five time-domain and five frequency-domain statistical features as mentioned in II-C, while the outputs comprise four classes: healthy, one, two, and three BRBs. These supervised ML algorithms have been extensively utilized in the condition monitoring of induction machines [1]–[4], [6]. In previous studies [2], [4], it has been demonstrated that not all supervised ML algorithms are suitable for condition monitoring of induction machines. This is justifiable as each supervised ML algorithm operates based on a different fundamental principle and may perform better for a specific problem or dataset. Therefore, our aim is to apply these six supervised ML algorithms and identify the most suitable one that can provide better accuracy with the augmented data.

III. Applications and Results

Figures 4 and 5 illustrate the interpolation of the mean and standard deviation of error values for the first time-domain feature TD_1 and frequency-domain feature FD_1 for healthy and one BRB conditions, respectively. Although interpolation has been performed for all five time and frequency-domain features, only one feature is presented here. Polynomial functions, namely linear, quadratic, and cubic, are fitted to estimate the mean ϵ^*_μ and standard deviation ϵ^*_σ of the error for all features based on maximum R-squared value. Linear and quadratic polynomial functions are fitted to the mean and standard deviation of the error for features TD_1 and FD_1 using four loading levels: 0%, 50%, 75%, and 100%. The mean

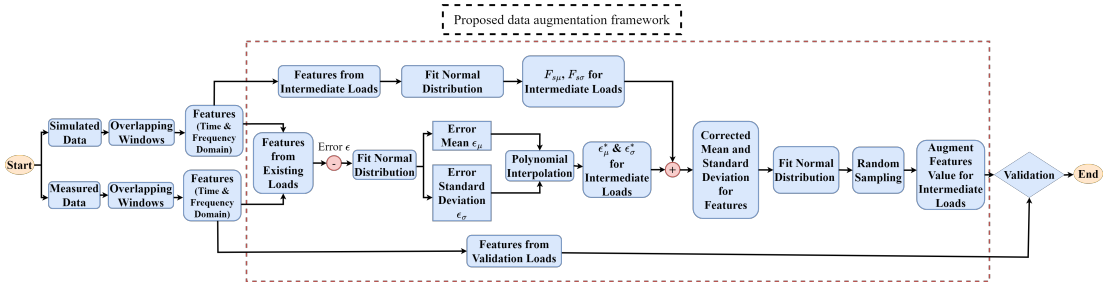

Fig. 2. Proposed data augmentation framework.

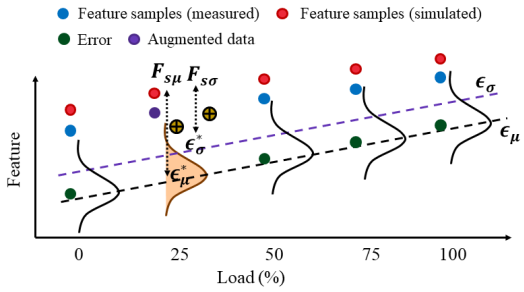

Fig. 3. A hypothetical illustration of a single feature value augmentation at 25% loading level. Symbol $\oplus$ indicates the addition of estimated mean ϵ^*_μ and standard deviation ϵ^*_σ of the error for each feature to mean $F_{s\mu}$ and standard deviation $F_{s\sigma}$ of features computed from simulated signal.

and standard deviation of the error for the unmeasured 25% loading level is estimated from the fitted curve. The estimated mean and standard deviation of the error at the 25% loading level is then validated against the actual values obtained from the difference of measured and simulated features values. It can be observed in Figs. 4 and 5 that the estimated mean and standard deviation of the error at the 25% loading level are closely aligned with the actual values computed from measured and simulated features values.

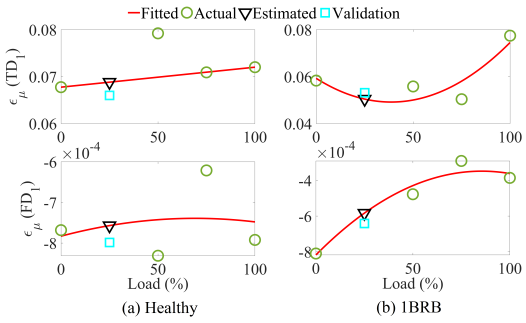

Fig. 4. Estimated mean of error for TD_1 and FD_1 features at 25% loading level for healthy and 1BRB conditions. Actual refers to the difference between measured and simulated features values.

Figure 6 illustrates the generated features values for TD_1 and FD_1 for healthy and one BRB conditions. The features computed from simulated data show a large deviation compared to the measured data, which can be observed in Fig. 6. On the contrary, the generated features values through the proposed augmentation technique remarkably minimize

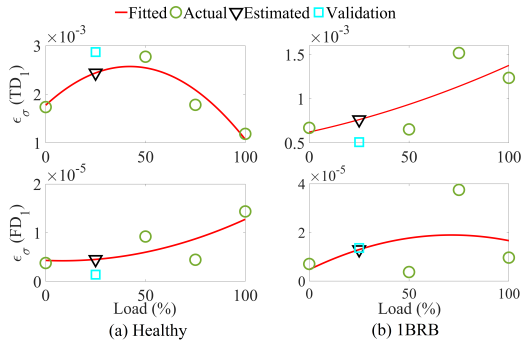

Fig. 5. Estimated standard deviation of error for TD_1 and FD_1 features at 25% loading level for healthy and 1BRB conditions. Actual refers to the difference between measured and simulated features values.

such deviation and exhibit excellent agreement with the actual features values calculated from the measured data.

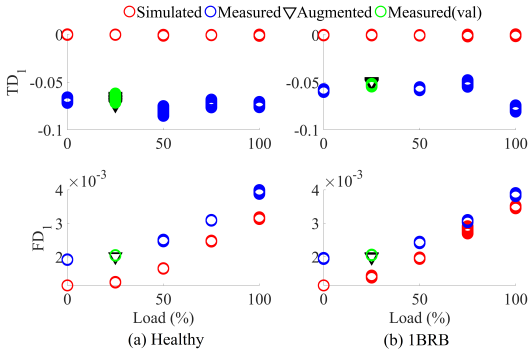

Fig. 6. Augmented values for TD_1 and FD_1 features at 25% loading level for healthy and 1BRB conditions.

We have examined four cases which are described in Table I. In Case-I, all five loading levels of the measured data are utilized for both training and testing. The data is randomly divided into training and testing sets, with a 70% to 30% ratio. In Case-II, ML models are trained using four loading levels by intentionally excluding one loading level to evaluate generalization. In Case-III, the missing loading level data is directly included in the training dataset from the MEC simulation. The performance of the proposed data augmentation technique is evaluated in Case-IV, where the excluded loading level data from the training dataset is compensated with generated data using the augmentation technique. It is

important to note that the test data always includes the specific loading level of the measured data that has been purposely dropped from the training dataset in Case-II, Case-III, and Case-IV. Hyperparameters of each ML algorithm are tuned using a random search with 5-fold cross-validation. The tuning is performed on the training dataset in Case-I, which includes all five loading levels of the measured data. The same tuned ML models are then applied to Case-II, Case-III, and Case-IV for the fair evaluation.

Figure 7 illustrates the average classification accuracy of ML algorithms on the test data for Case-I, Case-II, Case-III, and Case-IV. Most ML algorithms achieve 100% average accuracy, except for AdaBoost in Case-I. In the Fig. 7(a) and 7(b), the features values of 25% and 50% loading levels are excluded from the training dataset in Case-II, respectively. These missing loading levels features values are incorporated from the MEC simulation model in Case-III. In Case-IV, augmented data is used to replace the simulated data at the 25% and 50% loading levels in the training dataset. In Case-II, the average classification accuracy drops significantly due to the absence of data at the 25% and 50% loading levels, as seen in Fig. 7(a) and Fig. 7(b), highlighting a generalization problem. Moving on to Case-III, the addition of MEC simulated data enhances the accuracy of ML classifiers in Fig. 7(a). However, there is a notable decrease in accuracy in Fig. 7(b). This discrepancy can be attributed to the variations between simulated and measured data across different loading levels, as illustrated in Fig. 6, and larger differences can lead to poorer classification accuracy.

TABLE I
DESCRIPTION OF STUDIED FOUR CASES

Case study	Train data	Test data
Case-I	5600 data 70% of 5 loads/class (measured)	2400 data 30% of 5 loads/class (measured)
Case-II	6400 data 4 loads/class (measured)	1600 data 1 load/class (measured)
Case-III	6400 data + 1600 data 4 loads/class (measured) + 1 load/class (simulated)	1600 data 1 load/class (measured)
Case-IV	6400 data + 1600 data 4 loads/class (measured) + 1 load/class (augmented)	1600 data 1 load/class (measured)

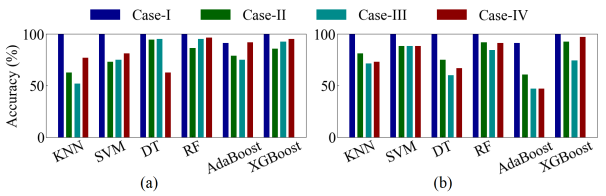

Fig. 7. Average classification accuracy on test data set for four studied cases; (a) 25% loading level data is studied (b) 50% loading level data is studied.

In Case-IV, the addition of augmented feature values for the 25% loading level notably enhances the classification accuracy of most ML classifiers, with the exception of DT, as illustrated in Fig. 7(a). However, when augmented data for the 50% loading level is introduced to the training dataset, it only improves the accuracy of the XGBoost classifier. The improvement in accuracy for other ML classifiers is less evident, as observed in Fig. 7(b). Notably, the XGBoost algorithm consistently demonstrates enhanced classification performance when augmented data for both 25% and 50% loading levels are integrated into the training dataset, compared to the scenarios presented in Case-II and Case-III. Nevertheless, the incorporation of augmented data in Case-IV does not yield the same level of accuracy observed in Case-I for the 25% and 50% loading levels. This discrepancy implies that a distinction still exists between the distribution of augmented and measured data. Consequently, it can be deduced that among the selected ML algorithms within this augmentation framework, the XGBoost algorithm proves to be the most suitable option.

IV. CONCLUSION

This article introduces a data augmentation technique for generating feature values at unmeasured loading levels. The generated feature values demonstrate a good agreement with measured data, addressing the generalization problem of the XGBoost algorithms. Our proposed method can be universally applied to generate feature values for various signals and faults in electrical machines. Future studies will investigate the robustness of this method against high-level noise, the inclusion of various loading levels in the interpolation process, and the extrapolation capacity.

ACKNOWLEDGMENT

This work was supported in part by the Academy of Finland consortium grant 330747.

REFERENCES

[1] M. Kang and J.-M. Kim, "Reliable fault diagnosis of multiple induction motor defects using a 2-d representation of shannon wavelets," *IEEE Transactions on Magnetics*, vol. 50, no. 10, pp. 1–13, 2014.

[2] M.-Q. Tran, M. Elsisi, K. Mahmoud, M.-K. Liu, M. Lehtonen, and M. M. Darwish, "Experimental setup for online fault diagnosis of induction machines via promising iot and machine learning: Towards industry 4.0 empowerment," *IEEE access*, vol. 9, pp. 115 429–115 441, 2021.

[3] A. N. Saberi, A. Belahcen, J. Sobra, and T. Vaimann, "Lightgbm-based fault diagnosis of rotating machinery under changing working conditions using modified recursive feature elimination," *IEEE Access*, vol. 10, pp. 81 910–81 925, 2022.

[4] M. Z. Ali, M. N. S. K. Shabbir, X. Liang, Y. Zhang, and T. Hu, "Machine learning-based fault diagnosis for single- and multi-faults in induction motors using measured stator currents and vibration signals," *IEEE Transactions on Industry Applications*, vol. 55, no. 3, pp. 2378–2391, 2019.

[5] S. M. K. Zaman and X. Liang, "An effective induction motor fault diagnosis approach using graph-based semi-supervised learning," *IEEE Access*, vol. 9, pp. 7471–7482, 2021.

[6] A. N. Saberi, S. Sandirasegaram, A. Belahcen, T. Vaimann, and J. Sobra, "Multi-sensor fault diagnosis of induction motors using random forests and support vector machine," in *2020 International Conference on Electrical Machines (ICEM)*, vol. 1. IEEE, 2020, pp. 1404–1410.

[7] L. Weili, X. Ying, S. Jiafeng, and L. Yingli, "Finite-element analysis of field distribution and characteristic performance of squirrel-cage induction motor with broken bars," *IEEE Transactions on Magnetics*, vol. 43, no. 4, pp. 1537–1540, 2007.

[8] M. Ojaghi, M. Sabouri, and J. Faiz, "Performance analysis of squirrel-cage induction motors under broken rotor bar and stator inter-turn fault conditions using analytical modeling," *IEEE Transactions on Magnetics*, vol. 54, no. 11, pp. 1–5, 2018.

[9] G. Y. Sizov, A. Sayed-Ahmed, C.-C. Yeh, and N. A. Demerdash, "Analysis and diagnostics of adjacent and nonadjacent broken-rotor-bar faults in squirrel-cage induction machines," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 11, pp. 4627–4641, 2009.