Deep reinforcement learning for optimizing operation and maintenance of energy systems equipped with PHM capabilities

Luca Pinciroli

Energy Department, Politecnico di Milano, Milan, Italy. E-mail: luca.pinciroli@polimi.it

Piero Baraldi

Energy Department, Politecnico di Milano, Milan, Italy. E-mail: piero.baraldi@polimi.it

Guido Ballabio

Aramis S.r.l., Milan, Italy. E-mail: guido.ballabio@aramis3d.com

Michele Compare

Aramis S.r.l., Milan, Italy. E-mail: michele.compare@aramis3d.com

Enrico Zio

Energy Department, Politecnico di Milano, Milan, Italy.
MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France.
Eminent Scholar, Department of Nuclear Engineering, College of Engineering, Kyung Hee University,
Republic of Korea. E-mail: enrico.zio@polimi.it

The Life Cycle Cost (LCC) of energy systems including Renewable Energy Sources (RES) strongly depends on the Operation and Maintenance (O&M) costs. Nowadays, many components of these energy systems are equipped with Prognostics & Health Management (PHM) capabilities, for estimating their current and future health states. This information is intended to be used for the optimization of O&M. It is an ambitious and challenging objective as the uncertain information brought by PHM must be combined with other factors influencing O&M, such as the limited availability of maintenance crews, the variability of energy demand and production, the long-time horizons of energy systems. In this work, we formalize the O&M optimization of RES-based energy systems equipped with PHM as a sequential decision problem over a long-time horizon and we solve it by Deep Reinforcement Learning (DRL). The proposed methodology is applied to a small wind farm. Strengths and weaknesses are analyzed by means of a comparison with state-of-the-art O&M policies.

Keywords: Energy Systems, Operation and Maintenance, Prognostics and Health Management, Optimization, Deep Reinforcement Learning.

1. Introduction

Prognostics and Health Management (PHM) uses condition monitoring data for estimating the equipment current health state and predicting its Remaining Useful Life (RUL). Several algorithms for RUL estimation have been developed (Simões et al. (2011); Liu et al. (2018)), and many successful applications are reported in literature (Hu et al. (2015); Kwon et al. (2016); Yang et al. (2016); Rigamonti et al. (2016); Xu et al. (2020)).

Predictive Maintenance (PdM) exploits PHM to set efficient, just-in-time and just-right maintenance interventions: in other words, provide the right part to the right place at the right time. This gives the opportunity of maximizing system availability, and minimizing the system Life Cycle Cost (LCC) and losses.

Although the advantages of PdM are intuitive, a clear understanding of the actual business opportunities of PdM for energy systems based on Renewable Energy Sources (RES) is still lacking. This is due to the following main reasons:

1) The prediction of the equipment RUL must consider its reciprocal relation with the Operation and Maintenance (O&M) decisions. That is, RUL predictions must consider the dynamic management of the equipment and its effects on the equipment future degradation evolution. For example, the RUL of the gearbox of a wind turbine is influenced by the applied loading conditions, which in turn depend on the wind conditions and the O&M decisions taken in time for optimal equipment usage and for responding to the power

Proceedings of the 30th European Safety and Reliability Conference and
the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Published by Research Publishing, Singapore.
ISBN: 978-981-14-8593-0; doi:10.3850/978-981-14-8593-0

demand. When predicting the RUL, these future conditions of equipment usage are generally assumed constant or behaving according to some known exogenous stochastic process, i.e., without considering the interwined relation of RUL with O&M decisions. This does not reflect reality and the RUL predictions that guide the O&M decisions are deemed to be incorrect (Bellani et al. (2020)) and can lead to sub-optimal decisions.

2) Solving the O&M optimization issue for RES-based energy systems with PHM requires considering several other factors, such as the long-time horizons, the limited availability of maintenance teams, the variability of energy demand and production, the long-time horizons that usually characterize industrial systems and the uncertainty related to all these pieces of information.

In this work the O&M management issue of RES-based energy systems equipped with PHM capabilities is formalized as a Sequential Decision Problem (SDP) over a long-time horizon, which is solved by Deep Reinforcement Learning (DRL) (Sutton and Barto (2018); Arulkumaran et al. (2017)).

Reinforcement Learning (RL) is a machine learninig framework in which a learning agent optimizes its behaviour by means of consecutive trial and error interactions with a white-box model of the system in order to find the optimal policy (Kaelbling et al. (1995); Grondman et al. (2012)), i.e. the function linking each system state to the action that maximizes a reward. RL has been shown to be suitable to solve complex decision-making problems in many fields (Li (2017)), including energy-related ones (Rocchetta et al. (2019)).

In principle, tabular RL algorithms allow finding the exact solution of SDPs (Sutton and Barto (2018)). However, in most cases, their computational cost is not compatible with realistic applications to complex systems. For this reason, we resort to DRL, using deep Artificial Neural Networks (ANNs) to find an approximate solution to the optimization problem.

The proposed framework is applied to a scaled-down case study concerning the optimization of the O&M strategy of a wind farm. The results are compared to state of the art and user-defined O&M policies.

The structure of the paper is as follows. In Section 2, we introduce the problem statement and in Section 3 we discuss its formulation as a SDP. In Section 4, details about the RL algorithm are provided. In Section 5, the case study concerning a wind farm is considered. Results are discussed in Section 6. Finally, conclusions are drawn in Section 7.

2. Problem Statement

We consider a plant of L identical Wind Turbines (WTs) equipped with PHM capabilities, estimating the RUL. The power production of each WT is strictly related to the enviromental conditions and the maintenance of the WT is performed by C maintenance crews.

For each WT $l \in \Lambda = \{1, \ldots, L\}$, the probability density function (pdf) of the failure time T_l is known. At any time t, we indicate the ground truth RUL of WT $l \in \Lambda$ as:

$$R_l^* = T_l - t \tag{1}$$

and the RUL estimate R_l by:

$$R_l = T_l - t + \epsilon_R \tag{2}$$

where $\epsilon_R \sim N(0, \sigma_R)$ is the Gaussian noise of the RUL estimate.

The production level P_l^* of the $l - th$ WT is a function of time, representing the proportion of produced power with respect to the maximum value. For simplicity, P_l^* is assumed independent from the component degradation level. At any time t, we predict the future production level P_l for the following J days with accuracy $\epsilon_P \sim N(0, \sigma_P)$:

$$P_l(t+j) = P_l^*(t+j) + \epsilon_P \qquad j = 1, ..., J \tag{3}$$

where $P_l^*(t+j)$ is the true production level.

A maintenance crew cr_c, with $c \in \Gamma = \{1, \ldots, C\}$, can reach the l-th WT and perform $i)$ Preventive Maintenance (PM), if the component is not failed, i.e. $R_l^* > 0$, or $ii)$ corrective maintenance (CM), if the component is failed, i.e. $R_l^* = 0$, or it can reach the depot, H, and wait up for the next decision time.

The downtimes of the WTs due to PM and CM actions, Π_{PM} and Π_{CM}, respectively, are random variables obeying probability density functions $f_{\Pi_{PM}}$ and $f_{\Pi_{CM}}$, respectively. The downtime of a PM action is expected to be shorter than that of a CM, as all the maintenance logistic support issues have already been addressed (Compare et al. (2018)).

The costs of the preventive and corrective maintenance actions on each WT are U_{PM} and U_{CM}, respectively.

The objective of the work is to define the optimal O&M policy, π^*, i.e. the optimal sequence of actions to be taken at every decision instant t in order to maximize the future rewards.

3. Problem Formulation

We formulate the problem as a SDP. Sections 3.1, 3.2 and 3.3 will define the state space, the actions and the rewards.

3.1. *State space*

The state at time t contains all the information retriveable from the system and its environment. It is defined by the vector $\mathbf{s}_t = [\mathbf{R}_t, \mathbf{P}_t, \mathbf{MT}_t, t]$, obtained appending vectors $\mathbf{R}_t = [R_1(t), ..., R_L(t)]$, $\mathbf{P}_t = [P_1(t+1), ..., P_L(t+1), P_1(t+2), ..., P_L(t+2), ..., P_1(t+J), ..., P_L(t+J)]$, the time needed to complete the current maintenance action on each WT, $\mathbf{MT}_t = [MT_1, .., MT_L]$ and the current time t. Then, $\mathbf{s}_t \in \mathbb{R}^{(2+J)\cdot L+1}$.

3.2. *Actions*

Every time step, a decision is taken about the next destination of each maintenance crew, chosen among the $L+1$ possible destinations. Namely, the learning agent returns as output a vector of C destinations, one per crew. The available O&M decisions are organized in vector $\mathbf{a}_t = [a_1, \ldots, a_{L+1}]$, where $a_l, l = 1, \ldots, L$ refers to setting the destination of the selected maintenance crew to component l, whereas the last action correspond to the decisions of sending the maintenance crew to the depot. If one of the L equipments is selected as the destination, the maintenace intervention (preventive or corrective) starts as soon as the crew reaches the equipment, whereas if the depot is selected as the crew destination it will start waiting for a new assignment as soon as it arrives at destination. When a maintenance operation starts, the corresponding component is stopped and its production level is set to 0.

3.3. *Rewards*

At every decision instant t, the decision maker receives a reward r_t defined as:

$$r_t = G_t - X_t \tag{4}$$

where $G_t = \sum_{l=1}^{L} C \times P_l^*(t)$ defines the revenues at time t, being K the maximum revenue per WT, whereas $X_t = \sum_{l=1}^{L} U_{PM} \times f_l^{PM}(t) + U_{CM} \times f_l^{CM}(t)$ defines the maintenace costs at time t, being $f_l^{PM}(t) = I_{R_l(t)>0}$ and $f_l^{CM}(t) = 1 - f_l^{PM}(t)$ two boolean variables representing the type of maintenance action performed on the $l-th$ component at time t.

4. Reinforcement Learning Algorithm

RL algorithms can be divided into thre groups: *value function*, *policy search* and *actor-critic* methods (Konda and Tsitsiklis (2000); Arulkumaran et al. (2017); Sutton and Barto (2018)). Value function methods learn the value of being in a particular state and, then, select the optimal action according to their estimated state values. A well known example of value function method is Deep Q-Networks (DQN) (Mnih et al. (2015)), in which a deep neural network is used to approximate the value function. Value function methods usually show slow convergence rate and fail on many simple problems (Schulman et al. (2017)).

Policy search methods directly look for the optimal policy by learning a parameterized policy through which optimal actions are selected. The update of the policy parameters can be performed by means of gradient-free methods, e.g., evolutionary algorithms, or gradient-based methods, e.g. REINFORCE algorithms (Williams (1992)).

Actor-Critic methods learn both the value funtion and the policy aiming at combining the strong points of value function and policy search methods (Konda and Tsitsiklis (2000)). Actor-Critic methods consist of two models: the critic which learns the value function and the actor which learns the policy updating the parameters in the direction suggested by the critic.

In this work, we have selected Proximal Policy Optimization (PPO) (Schulman et al. (2017)), an actor-critic method, as algorithm to define the optimal O&M policy. In PPO, an estimator of the gradient is computed by differentianting a surrogate objective defined as the minimum between an unclipped and a clipped version of a function of the reward (Schulman et al. (2017)). The minimum is used to define a lower, i.e. pessimistic, bound on the unclipped objective and the clipping is used to penalize too large policy update, avoiding second order approximations of a constraint, as in Trust Region Policy Optimization (TRPO) (Schulman et al. (2015)). Despite PPO relative simplicity of implementation, it has been shown to outperform many state-of-the-art approaches Schulman et al. (2017).

Since the state space is very large, it can be hard for the agent to find the optimal action to take in every state, starting from a random initialization of the neural network. For this reason we resort to behaviour cloning (Hester et al. (2017); Arulkumaran et al. (2017)), which consists in training the agent to reproduce a heuristic policy by means of supervised learning and, then, fine-tuning the agent using RL.

5. Case Study

We consider a wind farm composed of $L = 10$ identical WTs equipped with PHM capabilities. The failure time, T, of each WT is sampled from an exponential distribution with failure rate $\lambda_f = 6.58\ 10^{-3} \frac{1}{d}$, assuming λ_f to be equal to the mean value of the failure rates of different WT subsystems (Ozturk et al. (2018)). The predicted RUL, R, is estimated at each time according to Eq.(2), assuming $\sigma_R = 5d$.

To simulate the seasonality and the stochasticity of the wind velocity, respectively, we assume that the power production of each WT is the sum of a periodic behaviour and a noise:

$$P_l^*(t) = clip(0.65\, sin\left(\frac{2\pi t}{\tau}\right) + 0.5 + \epsilon_w,\ 0,\ 1) \quad (5)$$

where $\epsilon_w \sim N(0, \sigma_w)$, with $\sigma_w = 0.35$, and $clip(..., 0, 1)$ identifies the clipping operation between 0 and 1 and the sinusoidal component period, τ, is set equal to $24d$. Figure 1 shows a sampled trajectory of power production.

At every decision time t, the value of the predicted production for the next j days, with $j \in \{1, .., J\}$, is set according to Eq. 3, where $\sigma_P = 0.03$. We set the number of prediction days $J = 3d$.

The maintenance is managed by $C = 1$ maintenance crew. The maintenance times are sampled from exponential distributions with repair rate $\lambda_{PM} = 2.94\frac{1}{d}$ and $\lambda_{CM} = 1.83\frac{1}{d}$, for preventive and corrective maintenance respectively, setting λ_{PM} λ_{CM} equal to the mean value of the repair rates of different WT sub-systems (Carroll et al. (2015)).

Finally, the income $K = 96$, whereas the cost of PM and CM actions are $U_{PM} = 180$ and $U_{CM} = 2247$, all in arbitrary units.

6. Results

The RL optimized policy has been compared with several user-defined policies over 100 test episodes: i) a fully-random policy, ii) a corrective maintenance policy, iii) a scheduled maintenance policy in which the maintenance interventions are scheduled at regular intervals, iv) a predictive maintenance policy in which the maintenance interventions are performed only when the turbine RUL estimation is smaller than a user-defined threshold and v) a modified-predictive maintenance policy in which the information about both the turbine RUL and future power production is used for planning the maintenance interventions when both quantities are below user-defined thresholds. In all these policies, the thresholds have been set by optimizing the profit over 250 episodes using the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al. (2011)). The performance over 100 test episodes are reported in Table 1.

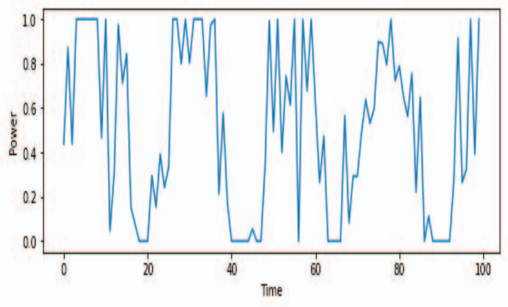

Fig. 1. Behaviour of the power production.

Table 1. Performance of the tested policies in terms of average profit over 100 test episodes.

Maintenance policy	Average profit
Random	78209.47 ± 22504.11
Corrective	323607.67 ± 19868.04
Scheduled	317332.58 ± 23452.75
Predictive	458919.72 ± 5677.08
Modified-predictive	460580.07 ± 7309.92
RL policy	462069.74 ± 4300.16

The RL optimal policy is characterized by a perfomance comparable to the user-defined heuristics. In particular, the RL policy provides better performance than the corrective and scheduled maintenance policies, which are the most commonly applied maintenance policies (Nilsson Westberg and Bertling Tjernberg (2007); Barberá et al. (2013); Asensio et al. (2015); Pattison et al. (2016); Chan and Mo (2017)), and performs very similarly to the predictive and modified-predictive policies, which exploit the information about the equipment health state.

Figure 2 shows the number of maintenance interventions performed by the predictive, modified-predictive and RL policies at different RULs, whereas Figure 3 shows the number of maintenance interventions performed by the same policies at different future power levels, normalized by the number of times each power level is verified. It can be noticed that even if the RL agent has been pre-trained by means of the predictive policy, the optimized RL agent has found a different optimal policy; in fact, the RL agent prefers to perform maintenance at larger RUL values with respect to the predictive and the modified predictive policies (Figure 2), and the predictive policy performs the same number of maintenance actions at every power level, the modified predictive policy performs many interventions at low power levels and performs few interventions at high power levels (no interventions at power higher that 0.9), whereas the RL optimal policy prefers to perform maintenance at low power levels but, differently from the modified-preventive policy, sometimes it performs maintenance when the power is equal to one, in order to avoid failures (Figure 3).

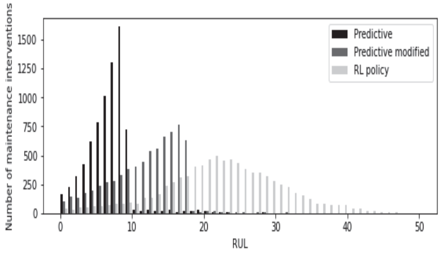

Fig. 2. Number of maintenance interventions at different RUL values.

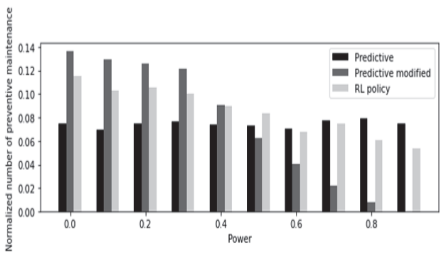

Fig. 3. Number of maintenance interventions at different power values.

7. Conclusions

In this paper we have performed a preliminary analysis to investigate the applicability of DRL to the optimization of the O&M policy of a RES-based energy system. For this, we have trained a neural network to choose the best action to be performed at each decision instance considering the available information about the system and its environment.

The proposed approach has been tested on a scaled-down wind farm and has been shown to provide an O&M policy, which outperforms state-of-the-art policies and behaves similarly to user-defined policies. Future work will consider a more realistic and complex environment. In particular, the wind turbine can be considered as a complex engineering system composed of several interacting components, each one characterized by different degradation behaviour, failure severity and impact on the power production, and the effect of the O&M decisions on the RUL estimation can be takan into account. Furthermore, the stochasticity of the wind velocity can be modeled to be more representative of a particul geographic location. Eventually, new environment parameters, such as the power demand, can be considered as variables and can be added to the state space in order to optimize the O&M policy.

Acknowledgement

The participation of Piero Baraldi and Enrico Zio has been funded by ”Smart maintenance of industrial plants and civil structures by 4.0 monitoring technologies and prognostic approaches - MAC4PRO”, sponsored by the call BRIC-2018 of the National Institute for Insurance against Accidents at Work – INAIL.

References

Arulkumaran, K., M. P. Deisenroth, M. Brundage, and A. A. Bharath (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine 34*(6), 26–38.

Asensio, S., J. M. Pinar Pérez, and F. P. García Márquez (2015, 01). *Economic Viability Study for Offshore Wind Turbines Maintenance Management*, Volume 362, pp. 235–244.

Barberá, L., A. Guerrero, A. Crespo Marquez, V. Gonzalez-Prida, A. J. Guillén Lopez, J. F. Gomez Fernandez, and A. Sola (2013, 01). State of the art of maintenance applied to wind turbines. Volume 33, pp. 931–936.

Bellani, L., M. Compare, P. Baraldi, and E. Zio (2020). Towards developing a novel framework for practical phm: a sequential decision problem solved by reinforcement learning and artificial neural networks. *International Journal of Prognostics and Health Management 10*.

Bergstra, J. S., R. Bardenet, Y. Bengio, and B. Kégl (2011). Algorithms for hyperparameter optimization. In *Advances in neural information processing systems*, pp. 2546–2554.

Carroll, J., A. Mcdonald, and D. Mcmillan (2015). Failure rate, repair time and unscheduled o&m cost analysis of offshore wind turbines. *Wind Energy 19*.

Chan, D. and J. Mo (2017). Life cycle reliability and maintenance analyses of wind turbines. *Energy Procedia 110*, 328 – 333. 1st International Conference on Energy and Power, ICEP2016, 14-16 December 2016, RMIT University, Melbourne, Australia.

Compare, M., L. Bellani, E. Cobelli, and E. Zio (2018). Reinforcement learning-based flow management of gas turbine parts under stochastic failures. *The International Journal of Advanced Manufacturing Technology 99*.

Grondman, I., L. Busoniu, G. A. D. Lopes, and R. Babuska (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42*(6), 1291–1307.

Hester, T., M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, A. Sendonaris, G. Dulac-Arnold, I. Osband, J. Agapiou, J. Leibo, and A. Gruslys (2017, 04). Learning from demonstrations for real world reinforcement learning.

Hu, Y., P. Baraldi, F. Di Maio, and E. Zio (2015). A particle filtering and kernel smoothing-based approach for new design component prognostics. *Reliability Engineering & System Safety 134*, 19 – 31.

Kaelbling, L. P., M. L. Littman, and A. W. Moore (1995). An introduction to reinforcement learning. In L. Steels (Ed.), *The Biology and Technology of Intelligent Autonomous Agents*, Berlin, Heidelberg, pp. 90–127. Springer Berlin Heidelberg.

Konda, V. R. and J. N. Tsitsiklis (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014.

Kwon, D., M. R. Hodkiewicz, J. Fan, T. Shibutani, and M. G. Pecht (2016). Iot-based prognostics and systems health management for industrial applications. *IEEE Access 4*, 3659–3670.

Li, Y. (2017). Deep reinforcement learning: An overview. *ArXiv abs/1701.07274*.

Liu, Z., Z. Jia, C. Vong, J. Han, C. Yan, and M. Pecht (2018). A patent analysis of prognostics and health management (phm) innovations for electrical systems. *IEEE Access 6*, 18088–18107.

Mnih, V., K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis (2015, 02). Human-level control through deep reinforcement learning. *Nature 518*, 529–33.

Nilsson Westberg, J. and L. Bertling Tjernberg (2007, 04). Maintenance management of wind power systems using condition monitoring systems—life cycle cost analysis for two case studies. *Energy Conversion, IEEE Transactions on 22*, 223 – 229.

Ozturk, S., V. Fthenakis, and S. Faulstich (2018). Failure modes, effects and criticality analysis for wind turbines considering climatic regions and comparing geared and direct drive wind turbines. *Energies 11*(9), 2317.

Pattison, D., M. D. S. Garcia, W. Xie, F. Quail, M. Revie, R. Whitfield, and I. J. Irvine (2016). Intelligent integrated maintenance for wind power generation. *Wind Energy 19*, 547–562.

Rigamonti, M. M., P. Baraldi, E. Zio, et al. (2016). Echo state network for the remaining useful life prediction of a turbofan engine. In *annual conference of the prognostics and health management society 2015*, pp. 255–270.

Rocchetta, R., L. Bellani, M. Compare, E. Zio, and E. Patelli (2019). A reinforcement learning framework for optimal operation and maintenance of power grids. *Applied Energy 241*, 291–301.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz (2015). Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Simões, J., C. Gomes, and M. Yasin (2011). A literature review of maintenance performance measurement: A conceptual framework and directions for future research. *Journal of Quality in Maintenance Engineering 17*, 116 – 137.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement Learning: An Introduction*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning 8*(3-4), 229–256.

Xu, M., P. Baraldi, S. Al-Dahidi, and E. Zio (2020). Fault prognostics by an ensemble of echo state networks in presence of event based measurements. *Engineering Applications of Artificial Intelligence 87*, 103346.

Yang, Z., P. Baraldi, and E. Zio (2016). A comparison between extreme learning machine and artificial neural network for remaining useful life prediction. In *2016 Prognostics and System Health Management Conference (PHM-Chengdu)*, pp. 1–7.