A physics-guided, ensemble learning-based prognostic for turbofan engine degradation.

Samantha Deokinanan

City University of New York, School of Professional Studies

Spring, 2021

Table of Content

Abstract

Degradation modeling and prediction of remaining useful life (RUL) are crucial to prognostics and health management of aircraft engines. Physics-based and data-driven models are often limited in their applicability to complex real-world domains due to (1) incompleteness of understanding all potential physics-of-failures and (2) limited representativeness of the training dataset for data-driven models. Hybrid models thus leverage the complementary strengths. While significant research has been conducted in prognostics, there has been little research reported on the RUL prediction using an ensemble learning method that combines prediction results from hybrid prognostics. The objective of this research is to introduce an ensemble learning framework that incorporates hybrid approaches to comprehend the degradation process and predict the RUL of an aircraft engine. More specifically, the first base-model is a hybrid model that implements an accelerated-degradation model, and the second base-model infers the thermodynamic and virtual health of the engine via recurrent neural networks. The proposed algorithm is evaluated using NASA's popular Modular Aero-Propulsion System Simulation dataset, which is the only publicly available dataset that provides health, control, and engine parameters of the turbofan engine. Experimental results showed that the physics-guided ensemble learning prognostic predicts the RUL of the aircraft engines with robustness while making use of a compact model.

Keyworks: ensemble learning, remaining useful life, health index, aircraft engine, prognostics and health management

Introduction

A gas turbine-based turbofan engine is considered a complex engineering system, see Figure 1. These are combustion engines that suck air into an inlet by fan blades. The air is steadily compressed as it travels along the engine body by the low-pressure compressor, then it is compressed further by the high-pressure compressor to increase its pressure and temperature. This highly compressed air is mixed with fuel and ignited by the engine combustor where the high- then low-pressure turbine ejects it with an immense thrust through the exhaust nozzle, propelling an aircraft forward. With this many activities occurring on a machine expected to operate 24/7, and transporting the lives of hundreds, scheduled maintenance is necessary.

The degradation of such a complex type of machinery is a natural process. Industries aim to decrease the costs associated with downtime and defective products and increase the reliability of their systems (Li, Z., Goebel, K., *et al*, 2019). In the field of aircraft maintenance, global maintenance, repair, and overhaul spend on commercial aircraft engines in 2016 was valued at $27 billion (Lee, S. *et al*, 2008). Therefore, Prognostics and Health Management are often armed with state-of-the-art methods to estimate the health status of a piece of equipment. Such maintenance can detect pending failures and in turn, lead to immediate prevention. However, evaluating the health status of an engine is not a straightforward task. Attempts to resolving this issue have included combining the use of historical data, ad hoc defined health status, statistical inference methods, and engineering approaches to estimate the remaining useful life (RUL) of the engine, the predicted time of failure for any engineering system.

The existing literature on RUL prediction for aircraft engines can be classified into three categories: physics-based, data-driven, and hybrid prognostics. Physics-based prognostics rely on a critical understanding of the physical system to accurately capture the system behavior and degradation progression (Hu, C., *et al*, 2012; Laio, L & Köttig, F, 2014). But often certain knowledge is unavailable in practice and disable models from determining mechanism failure. To address this issue, machine learning (ML) algorithms have been comprehensively utilized to obtain more precise and reliable prognostic results (Li, Z., Goebel, K., *et al*, 2019). These models can learn hidden correlations, the health of the system, and other underlying trends using data retrieved from multiple sensors that simultaneously monitor the degradation process of the engine. However, a vast among of data is required for training and it is difficult to determine which and what type of learning algorithm should be selected. Moreover, these models are limited by the representativeness of the training datasets. This led to the integration of multiple models to fill in the missing pieces. The key in doing this is to further leverage their complementary strengths. The expectation is that integrated physics-ML models can better capture the dynamics of scientific systems together (Willard, J., *et al*, 2020). The combination can be multiple ML algorithms in an ensemble prognostic, or a combination of physics-based prognostics with ML approaches in a hybrid/physics-guided prognostic. By combining the advantages of models, these types of prognostics have shown promising performance in estimating the RUL of aircraft engines than single model-based approaches (Laio, L & Köttig, F, 2014; Li, Z., Goebel, K., *et al*, 2019).

Problem Statement

The general problem is that there is no standard way of predicting RUL resulting in numerous methodologies aiming for robustness, accuracy and reliability (Li, Z., Goebel, K., *et al*,, 2019). It is challenging to determine which algorithm should be selected among the various competing machine learning algorithms. There are also several methods that can be used to address the problem of dynamical model calibration on well-known physical principles (Willard, J., *et al*, 2020).

Starting from an unknown initial health condition, the engine system's components experience normal, linear degradation until point in time where an abnormal condition arises leading to an eventual time of failure (Chao, M. A., *et al,* 2020). Moreover, the time to reach same level of degradation by engines of the same specifications is often different (Malhotra, P., *et al*, 2016). The specific problem is that while studies have combined the advantages of differently skilled models, there is still an inadequate exploration of ensemble learning-based prognostics that particularly implement calibration on well-known physical principles. To overcome this shortcoming, this project proposes an ensemble approach that employs calibrated physics-based performance models that consider the stochastic properties of the degradation process with machine learning architectures that considers the temporal dependencies of each sensor measurement. To help showcase the performance, this project utilized NASA's C-MAPSS dataset containing 160,359 rows of 21 sensor readings from 709 engine units. It is the benchmark dataset that is widely used in engine health management and prognostics analysis.

Literature Review

System Prognostics & Damage Propagation Modeling

Sensors signals contain a rich amount of information about the degradation status of an engine. To improve the predictions beyond that of state-of-the-art physical models, most prognostics are data-driven that meticulously model the data to directly predict RUL, often rendering physics-models as obsolete. This is known as general path modeling, where the variability within the data is captured using ML algorithms. Some of the algorithm these data-driven models utilize include random forests (Wu, D., *et al*, 2017), support vector regression (Khelif, R., *et al*, 2017), relevance vector machine (Zheng, Y., *et al*, 2014), and more to achieve better predictive performance. Also, with more advanced and computational power, there have been many newer algorithms such as evolutionary computation (Laredo, D., 2019), long short-term memory (Lan, G., *et al*, 2019), reduced kernel recursive least squares algorithm (Zhou, H., *et al*, 2017), and more.

On the contrary, most physics-guided models would firstly estimate a discrete health state of an engine from the sensory signal, namely the health index (HI), and the trend over time is referred to as the health index curve (Kim, M., 2019, Malhotra, P., *et* al, 2016). Each of these margins is normalized to the range (0,1), where one represents a healthy state and zero represents a faulty state of the health index of the engine. These models would implement various techniques to capture the important pattern in the degradation time series subsequences, thus obtaining information that can be used to differentiate between normal and degraded regions in the data. Once the HI curve is constructed, hybrid models then use it to estimate the RUL by comparing it to the trends of the failed instance, i.e. a curve matching method. Recent literature has developed multiple approaches that can solve the extrapolation of HI used for prediction of RUL. Liu and Huang (2016) presented a construct where data-level fusion is used to correlate the HI with sensor signals. Nieto, G. P. (2016) introduced a hybrid particle swarm optimization support vector regression-based model to predict the RUL of aircraft engines. Yu, J. (2017) reported the combination of logistic regression with particle filtering techniques for engine health status and prediction. In these cases, the evolution of a sensor signal is characterized to capture the stochastic nature of the degradation process.

Characterizing the health state of an engine has proven its advantages in practice (Malhotra, P., *et al*, 2016). For instance, it is often regarded as another sensor signal but with more information. It is considered a more real-time measure of the degradation process, and with easy interpretation, it allows for a faster decision on maintenance. However, despite the promising prognostic performance of these models, they all remain heuristic and still have challenges comprehending the degradation status because it is understanding all potential physics-of-failures and their interactions for a complex system is almost impossible. And so, there is a continuation of efforts to reduce the errors in estimations.

Ensemble Learning-Based Predictive Modeling

Ensemble learning methods are meta-algorithms that aggregate predictions from multiple algorithms, i.e. base learners, into a single higher-order model in order to improve predictive performance (Li, Z., Goebel, K., et al,). The most simplistic set-up for an ensemble is using the predictions from each base learner and taking the arithmetic average or weighted average to obtain the new predictions. Other advanced techniques that are commonly used include boosting, bagging, and stacking. Boosting construct base learners sequentially and then reduce the bias of the combined base learners (e.g. AdaBoost), while bagging builds multiple base learners independently and then averages their predictions or takes a weighted sum of their predictions (e.g. Random Forest). Whereas, in stacking, unlike in bagging, the base models or level-1 models are typically different and fit on the same dataset instead of samples, and unlike in boosting, stacking has a super-model or meta-model or level-1 model that is used to learn how to best combine the predictions from the contributing models (Breiman, L., 1996).

It should be noted that while its predictive abilities typically outperform any of the constituent learning algorithms alone, ensemble models do not warrant results are an enhancement in all circumstances. Attaining a performance improvement depends on the complexity of the assignment and whether it is adequately reproduced by the training data. Moreover, the errors in the predictions obtained from the base learners must be complex enough that there is more to determine by combining predictions. This suggests the dependence upon the choice of base models and whether they are sufficiently skillful and uncorrelated enough in their predictions (or errors) (Breiman, L., 1996).

In the field of aircraft maintenance, previous studies have demonstrated the effectiveness of ensemble learning for estimating engine remaining useful life. More specifically, Li, Z., Wu, D., et al (2019) developed an ensemble learning-based method that considers the effects of time-dependent degradation; Li, Z., Goebel, K., et al, (2019) method employed several base learning algorithms with sequential quadratic optimization and particle swarm optimization; Zhang, C., et al, (2016) introduced a multi-objective deep belief networks ensemble method; and Hu, C., et al, (2012) approach employed the multiple base-learner with a weighted-sum formulation in a k-fold cross validation schema. Wen, L., et al, (2019) implemented residual Convolution Neural Networks with k-fold cross-validation. Peel, L. (2008) used the Kalman filter to combine prediction results produced by different artificial neural networks. Each of these experimental results has shown that their respective ensemble method outperforms constituent models.

While much research has been conducted around Prognostics and Health Management, there is still a need for reliable prognostics that can determine the RUL of a system. Most of the aforementioned ensemble learning prognostics are data-driven. Following this exhaustive literature review, no previous research has investigated an approach that integrates physics-guided prognostic which constructs the health status of aircraft engines into an ensemble as a method of determining the remaining useful life. Therefore, to fill the gap, this research presents an ensemble learning framework that incorporates the more recent hybrid approaches as a means to comprehend the system health degradation and predict the remaining useful life. More specifically, the first base-model is a hybrid model that constructs the HI curve using an accelerated-degradation model, and the second base-model infers the thermodynamic and virtual health of the engine via deep learning architectures.

Methodology

Proposed Physics-Ensemble Framework

A generic computation framework of the ensemble learning-based prognostic is illustrated in Figure 2. This ensemble machine learning algorithm forms a stacked generalization model. The architecture of a stacking model involves two or more base models, i.e., the level-0 models, that fit and make predictions on the training data. These predictions are then compiled, and a meta-model, i.e., level-1 model, combines them to produce the final predictions (Breiman, L., 1996). The proposed ensemble model must be robust enough so that it can accurately predict the RUL. Therefore, stacking is considered an appropriate method since the two models presented next are considered complex and have different skills to evaluate the data set. This would ensure that the base model predictions (or the errors in predictions) are highly uncorrelated. The meta-learner, on the other hand, is often simple, and in this ensemble, a linear regression is used, see Table 1.

Training Base Models

Accelerated-Degradation Life Testing

To make the prognostics result more precise and reliable, the first hybrid model uses the compacted data and estimates the health index curve by combining the Arrhenius model from accelerated life testing and the Wiener process to create an accelerated-degradation model for a turbofan engine (Ghorbani, S. *et al*, 2020). By implementing the Arrhenius model, chemical reaction wear is accounted for. While Wiener process, also known as the Brownian Motion, allows the model to capture the stochastic properties of the degradation process. With the health index curve obtained from the accelerated-degradation model, the health index is calculated via a linear regression and the result is used as a covariate in an inverse Gaussian distribution to predict the RUL of the engine.

Thermodynamic Performance Model using RNN

In contrast to the hybrid architectures cited above, the second model, as proposed by Chao, M. A., *et al*, (2020), leverages inferred unobserved process properties and parameters via virtual sensors and combines it with the sensor readings from the condition monitoring data to estimates of the health condition, see Table 2. This hybrid-based model uses a deep neural network to inform the system’s health conditions. Specifically, a multilayered Recurrent Neutral Network (RNN) with dropout (i.e. regularization) is trained in an unsupervised manner to minimize the loss function given by the squared reconstruction error. The final hidden state is then used to estimate the health index at any point of time by comparing recent sensor history with periods of normal behavior. Note that mechanical friction is often nonlinear. To address this, a long short-term memory network is added. By leveraging RNN instead of the Convolutional Neural Networks (CNN) as demonstrated by Chao, M. A., *et al*, (2020), the health of the system is learned via a temporal model that tries to capture the complexity and instantaneous dependencies between sensor data (Gugulothu, N., *et al*, 2017). As a result, a compacted data set is not required for this model. Lastly, it is assumed that the target health index curve follows a general exponential degradation trend based on the last cycle corresponding to the end-of-life, and as such the health index is determined, and the RUL is predicted.

NASA C-MAPSS Dataset

Developed by NASA, the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) is a tool for MATLAB and Simulink environment which models a nonlinear, dynamic, and component-level engine control system for commercial, high-bypass, dual-spool turbofan engines. The designs include five rotating components which can be assessed: fan, low-pressure compressor (LPC), high-pressure compressor (HPC), high-pressure turbine (HPT), and low-pressure turbine (LPT). The aircraft turbofan engine diagram simulated by C-MAPSS is shown in Figure 1. The data provided is therefore from a high-fidelity system-level engine designed to simulate realistic flight conditions for a small fleet of aircraft engines and presents trajectories for nominal and fault engine degradation over a series of flights. This simulation notes that the damage accumulation is different for each engine and is not directly measurable based on the duration or situation of flight (Saxena. A., *et al*, 2008).

The C-MAPSS datasets contain 4 training files, 4 testing files, and 4 RUL files. Each time series is from a different aircraft engine and starts with various degrees of initial wear and manufacturing variation. The data subset is composed of corresponding engine unit number, operational cycle, three operational settings: altitude (0-42K ft.), Mach number (0-0.84), and throttle resolver angle (20-100), and multivariate temporal data coming from 21 sensors, see Table 2. At the start of the engine, the operational status is considered healthy and degrades as time progresses until a failure occurs. For the training sets, this degradation increases in significance until failure, while for the test sets, the engine stops at some point in time before failure. The RUL files include the actual RUL of the testing units, and these are used for evaluating the estimated RULs. Lastly, there are two failure modes in these datasets. One of the failure modes is due to HPC degradation, while the other one is owing to both HPC and fan degradation, see Table 3.

Data Preparation

Data Normalization

The C-MAPSS data set is composed of a simplistic non-trivial, mixture distribution of noise characteristic to simulate the main sources of true noise from real data, typically from manufacturing and assembly variation, process noise, measurement noise, etc. (Saxena. A., *et al*, 2008). When comparing features and the range values, there is a large difference, see Figure 3. Therefore, to de-noise the dataset, normalization is conducted to enable unbiased contribution from the output of each sensor. This results in a new dataset $\sim N(0, 1)$ for each sensor reading, as shown in Figure 4.

Feature Selection & Extraction

The data preprocessing is based on previous studies such as Ghorbani, *et al*. (2020), Ellefsen, *et al.* (2019), and Li, Z., *et al*,(2019) where it was analyzed that for subset FD001 and FD003, some of the 21 sensors do not present much variance or convey redundant information throughout the entire engine life cycle. The important variables were further confirmed using a Random Forest model, see Figure 5. Therefore, the unnecessary are discarded, leaving only 14 of the 21 sensor measurements: {2, 3, 4, 7, 8, 9, 11, 12, 13, 14, 15, 17, 20, 21}, and with only a single operating condition, the three operational settings are excluded. However, subset FD002 and FD004 are under six operating conditions, making it more difficult for the prognosis algorithm to detect the degradation behavior of the engine. Moreover, there are high correlations among the features, see Figure 6. As a result of the employing dimensionality reduction method, the Uniform Manifold Approximation and Projection (UMAP) to confronting nonlinear systems, a compacted sample with a lower dimension (n = 7) of extracted features was obtained. The advantages are that UMAP is much faster, preserves the global structure, use little computation power than other nonlinear methods such as t-SNE and Kernel PCA which is typically performed in data-driven modeling investigating RUL (McInnes, L. *et al*. 2018). Lastly, the C-MAPSS dataset is limited to the features the practitioner elects as outputs. To capture as much information about the conditions of the system engine, variables that can be derived based on those available should be derived. In this case, only the fuel flow (Wf) was determined using the total pressure of the HPC outlet and phi ratio.

Performance Evaluation

Prognostic health management aims to avoid failures; therefore, it is desirable to predict early failures as compared to predicting late since it may result in more severe consequences. The root-mean-square error (*RMSE*) and NASA's scoring function (s) are mainly used to quantitatively evaluate the performance of models. RMSE measures prediction accuracy and penalizes both early and late predictions. The scoring function is asymmetric and penalizes the estimated RUL value that is larger than the actual RUL value more heavily than early predictions. The formulation of these evaluation metrics are as follows:

$$RMSE = \sqrt{\frac{1}{m_*}\sum_{j=1}^{m_*}(\Delta^{(j)})^2} \qquad\qquad s = \sum_{j=1}^{m_*} \exp\left(\alpha\left|\Delta^{(j)}\right|\right)$$

where m_* denotes the total number of test data samples, $\Delta^{(j)}$ is the difference between the estimated and the real RUL of the j^{th} sample (i.e., $y^{(j)} - \hat{y}^{(j)}$) and α is $\frac{1}{13}$ if RUL is under-estimated and $\frac{1}{10}$ otherwise.

Experimental Results & Discussion

This research was performed on Google Collaboratory (Colab, for short). It is a platform that allows for Python execution on Google's could servers, leveraging GPUs and TPUs via a simple browser. Moreover, Colab is equipped with an extensive ML library that requires no system configurations. In this case, the Keras libraries, with TensorFlow backend, is a powerful open-source Python library for developing and evaluating deep learning models such as RNN.

The data generated from C-MAPSS are used to simulate the proposed model for the turbofan engine. With an average reduction of 42% in the RMSE, the ensemble model outperforms each of the constituent models. It can be seen from Table 4 that RMSE for the ensemble is the smallest amongst each sub-test dataset. Similarly, the S-scores are also the smallest suggesting that the prediction accuracy of the ensemble has fewer errors than the constituent models. With the S-score penalizing over-estimation more than under-estimation, the reduction in the S-score suggests that the method handles RUL over-estimation more efficiently. By taking a closer look at the estimations, it appears that as the true RUL of an engine increases, predictions increase in error. This suggests that the model can infer the engine health status at the initial stages of an engine's life, but less likely to reconstruct the health as the number of cycles increases, see Figure 7. The test data with one operational condition, i.e. FD001 and FD003, had the lower RMSEs, of which FD001 was the smallest. This suggests that under one operation condition, the model performed well, but error increases when faults are due to more components, i.e. FD003 having two fault modes. Moreover, FD002 and FD004 are under six operation conditions, with FD004 have two fault modes, and FD004 had the largest RMSE among the 4 sets. This suggests that the model has difficulties learning the complexity and instantaneous dependencies between sensor data when more operation conditions are present.

Comparison with the Literature

A key objective in the research is to leverage physics-ML models' complementary strengths. The accelerated-degradation model has proven its skills in assessing the chemical reaction wear and models the stochastic properties of the degradation process. While the thermodynamic performance model using RNN with LSTM is trained to assess the nonlinear dynamics of mechanical wear via a time series on condition monitoring sensors and virtual sensors. Given enough data, this model is expected to better capture the dynamics of scientific systems with the advantages of understanding the underlying physical processes. Finally, according to the predicted health index, a threshold can be established to estimate the RUL.

The performance of the proposed physics-guided ensemble learning model is compared against other ensemble models in the literature pertaining to aircraft prognostics and health maintenance. Before a model is selected to be compared, three criteria needed to be fulfilled. These include: 1) that the model must be an ensemble learning algorithm, 2) must use the standard data preparation guidelines for the C-MAPSS datasets as a benchmarking tool, and 3) must report at least the RMSE or a metric in which the RMSE can be derived (e.g. MSE).

The sub-dataset FD004 is used to compare the proposed ensemble model with other ensemble models from the literature due to the complexity inherent in its six operating conditions and two fault modes. As shown in Table 5, the physics-guided ensemble learning model does not outperform most of the ensemble models in the literature pertaining to aircraft prognostics and health maintenance. When comparing the proposed model to the two models that are very close in RSME, namely Residual CNN k-fold ensemble (Wen, L., *et al*, 2019) and Multi-objective

Deep Belief Networks ensemble (Zhang, C., *et al*, 2016), it can be stated that the proposed model gained most of its efficiency via the RNN training from the second base-model. There is no further employment of optimization techniques in the ensemble schematic, unlike the better performing models. Thus, the novelty of this methodology is that the proposed ensemble is simple to understand and implement, and lightweight.

Conclusion

In this project, an ensemble learning-based prognostic approach for predicting the RUL of an airplane engine with calibrated physics-based performance models and machine learning architectures was introduced. Each model infers the system health by modeling the degradation using physical principles. Subsequently, this information is combined with sensor readings and used as input to produce a reliable hybrid prognostics model. The ensemble thus accounts for the stochastic properties of the degradation process by combining the Arrhenius model from accelerated life testing with the Wiener process, and the nonlinearity of mechanical wear processes and temporal dependencies of each sensor measurement is learned via Recurrent Neural Network with Long-Short Term Memory networks. By combining these models' strengths, the ensemble approach achieves better accuracy in RUL predictions compared to any sole member algorithm. And so, the main contribution of this project is to establish that not only can hybrid models reconstruct a health index, but when combined, they can provide substantial insights about the system when multiple principles and machine learning algorithms are used to minimize the uncertainties involved in prognostics.

There are a few important topics for future research. Firstly, the proposed method should be extended to incorporate optimization methodologies to support training. A suggestion on such techniques includes weight optimization even adaptive weights. This can enable the ensemble to further evaluate the final predictions. Secondly, further studies are needed to understand how the proposed method can efficiently solve cases with more than one failure mode or operation condition, as it does for a single failure mode or operation condition. Additionally, the performance of adding more hybrid models with different skillsets in estimating the health index of an engine system can be further incorporated into the ensemble and evaluated for robustness.

As final remarks, it should be emphasized that there is no right way to predict RUL. For each model, whether data-driven, physics guided, or hybrid, understanding and recording all potential physics-of-failures and inferring their interactions for a complex system is almost impossible. Nonetheless, this experimental study has shown that a simple and lightweight ensemble can be employed to estimate the RUL. The proposed methodology does not conflict with the existing prognostics. And lastly, this project further shows that the construction of the health index curve by physics-guided models can be directly used for degradation modeling and prognostics and enhances performance for gas turbine engine prognostics.

Works Cited

1. Breiman, L. (1996). Stacked regressions. Machine learning, 24(1), 49-64.
2. Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2020). Fusing physics-based and deep learning models for prognostics. arXiv preprint arXiv:2003.00732.
3. Ellefsen, A. L., Bjørlykhaug, E., Æsøy, V., Ushakov, S., & Zhang, H. (2019). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. Reliability Engineering, 183, 240–251.
4. Ghorbani, S., & Salahshoor, K. (2020). Estimating Remaining Useful Life of Turbofan Engine Using Data-Level Fusion and Feature-Level Fusion. Journal of Failure Analysis & Prevention, 20(1), 323–332.
5. Gugulothu, N., Tv, V., Malhotra, P., Vig, L., Agarwal, P., & Shroff, G. (2017). Predicting remaining useful life using time series embeddings based on recurrent neural networks. arXiv preprint arXiv:1709.01073.
6. Hu, C., Youn, B. D., Wang, P., & Taek Yoon, J. (2012). Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. Reliability Engineering & System Safety, 103, 120–135.
7. Khelif, R., Chebel-Morello, B., Malinowski, S., Laajili, E., Fnaiech, F., & Zerhouni, N. (2016). Direct remaining useful life estimation based on support vector regression. IEEE Transactions on industrial electronics, 64(3), 2276-2285.
8. Kim, M. (2019). A Generic Health Index Approach for Multisensor Degradation Modeling and Sensor Selection. IEEE Transactions on Automation Science, 16(3), 1426–1437.
9. Lan, G., Li, Q., & Cheng, N. (2018, August). Remaining Useful Life Estimation of Turbofan Engine Using LSTM Neural Networks. In 2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC) (pp. 1-5). IEEE.
10. Laredo, D. (2019). A neural network-evolutionary computational framework for remaining useful life estimation of mechanical systems. Neural Networks, 116, 178–187.
11. Lee, S., Ma, Y. S., Thimm, G., & Verstraeten, J. (2008). Product Lifecycle Management in Aviation Maintenance, Repair and Overhaul. Comput. Ind., 59(2–3), 296–303.
12. Li, Z., Wu, D., Hu, C., & Terpenny, J. (2019). An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction. Reliability Engineering & System Safety, 184, 110-122.
13. Li, Z., Goebel, K., & Wu, D. (2019). Degradation Modeling and Remaining Useful Life Prediction of Aircraft Engines Using Ensemble Learning. Journal of Engineering for Gas Turbines & Power, 141(4), 1–10.
14. Liao, L., Köttig, F. (2014). Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. IEEE Trans Reliab, 63(1), 191–207.
15. Liu, K., and Huang, S. (2016). Integration of Data Fusion Methodology and Degradation Modeling Process to Improve Prognostics. IEEE Transactions on Automation Science and Engineering, 13(1), 344–354.
16. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
17. Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). Multi-sensor prognostics using an unsupervised health index based on LSTM encoder-decoder. arXiv preprint arXiv:1608.06154.

18. Nieto, P. G., Garcia-Gonzalo, E., Lasheras, F. S., and de Cos Juez, F. J., (2016). Hybrid PSO–SVM-Based Method for Forecasting of the Remaining Useful Life for Aircraft Engines and Evaluation of Its Reliability," Reliab. Eng. Syst. Saf., 138, 219–231.
19. Peel, L. (2008). Data driven prognostics using a Kalman filter ensemble of neural network models. In 2008 international conference on prognostics and health management (1-6). IEEE.
20. Saxena, A., & Goebel, K. (2008). Turbofan Engine Degradation Simulation Data Set, NASA Ames Prognostics Data Repository. [Online]. Available: http://ti.arc.nasa.gov/project/prognostic-data-repository
21. Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation. In Proceedings of the 1st International Conference on Prognostics and Health Management (PHM08), Denver, CO, USA, 6–9
22. Wen, L., Dong, Y., & Gao, L. (2019). A new ensemble residual convolutional neural network for remaining useful life estimation. Mathematical biosciences and engineering : MBE, 16(2), 862–880. https://doi.org/10.3934/mbe.2019040
23. Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. arXiv:2003.04919
24. Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. Journal of Manufacturing Science and Engineering, 139(7).
25. Yu, J., (2017). Aircraft Engine Health Prognostics Based on Logistic Regression With Penalization Regularization and State-Space-Based Degradation Framework, Aerosp. Sci. Technol., 68, 345–361.
26. Zhang, C., Lim, P., Qin, A., and Tan, K. C., (2017) Multiobjective Deep Belief Networks Ensemble for Remaining Useful Life Estimation in Prognostics. IEEE Trans. Neural Networks Learn. Syst., 28(10), 2306–2318.
27. Zheng, Y., Wu, L., Li, X., & Yin, C. (2014). A relevance vector machine-based approach for remaining useful life prediction of power MOSFETs. In 2014 Prognostics and System Health Management Conference (PHM-2014 Hunan) (pp. 642-646). IEEE.
28. Zhou, H., Huang, J., & Lu, F. (2017). Reduced kernel recursive least squares algorithm for aero-engine degradation prediction. Mechanical Systems & Signal Processing, 95, 446–467.

Appendix A: Tables

Table 1: Stacked generalization model of the proposed physics-guided ensemble learning progonostic

Level-0 Models	• Accelerated-Degradation Life Testing • Thermodynamic Performance Model using RNN
Level-1 Model	• Linear Regression

Table 2: Description of the C-MAPSS Data set

Sensor No.	Symbol	Description	
-	-	Engine Unit No.	
-	-	Time in cycles	
-	alt	Altitude (ft)	condition monitoring signals
-	XM	Mach number	
-	TRA	Throttle-resolver angle (%)	
1	T2	Total temperature at fan inlet (^{o}R)	
2	T24	Total temperature at LPC outlet (^{o}R)	
3	T30	Total temperature at HPC outlet (^{o}R)	
4	T50	Total temperature at LPT outlet (^{o}R)	
5	P2	Pressure at fan inlet (psia)	
6	P15	Total pressure in bypass-duct (psia)	
7	P30	Total pressure at HPC outlet (psia)	
8	Nf	Physical fan speed (rpm)	
9	Nc	Physical core speed (rpm)	
11	Ps30	Static pressure at HPC outlet (psia)	
13	NRf	Corrected fan speed (rpm)	
14	NRc	Corrected core speed (rpm)	
-	Wf	Fuel flow	
10	EPR	Engine pressure ratio (P50/P2)	
12	Phi	Ratio of fuel flow to Ps30 (pps/psi)	
18	Nf_dmd	Demanded fan speed (rpm)	
19	PCNfR_dmd	Demanded corrected fan speed (rpm)	
15	BPR	Bypass ratio	virtual sensors
16	farB	Burner fuel-air ratio	
17	htBleed	Bleed Enthalpy	
20	W31	HPT coolant bleed (lbm/s)	
21	W32	LPT coolant bleed (lbm/s)	

Table 3: C-MAPSS Datasets

Datasets	Fault Modes	Operation Conditions	Train Units	Total Train Cycles	Test Units	Total Test Cycles
FD001	1 (HPC)	1	100	20631	100	13096
FD002	1 (HPC)	6	260	53759	259	33991
FD003	2 (HPC + Fan)	1	100	24720	100	16596
FD004	2 (HPC + Fan)	6	249	61249	248	41214

Table 4: Model performance on the CMAPSS data sets

Prognostic Approach	FD001		FD002		FD003		FD004	
	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score
Model #1: Accelerated-Degradation Model*	37.62	140.99	70.59	416.37	47.03	322.28	72.81	385.56
Model #2: Thermodynamic RNN*	33.17	37.72	38.29	150.23	33.74	60.90	37.49	137.42
Proposed physics-guided Ensemble Model	**28.54**	**34.10**	**29.26**	**88.91**	**28.91**	**45.46**	**29.41**	**69.51**

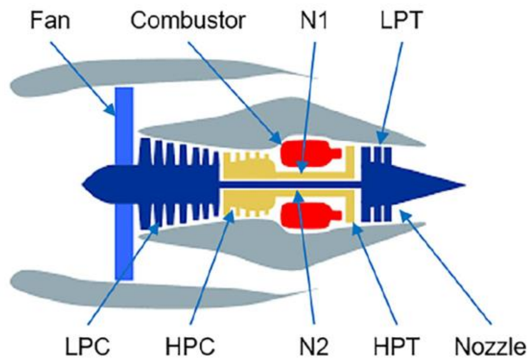
* - algorithms used in the proposed physics-guided Ensemble Model

Table 5: Performance comparison with the ensemble learning methods and other existing ensemble model in the literature on the CMAPSS FD004 data set

Prognostic Approach	FD004
	RMSE
Optimized, degradation-dependent weights ensemble (Li, Z., Wu, D., *et al*, 2019)	24.14
Residual CNN k-fold ensemble (Wen, L., *et al*, 2019)	28.56
Multi-objective Deep Belief Networks ensemble (Zhang, C., *et al*, 2016)	28.66
Proposed physics-guided ensemble model	**29.41**
Multi-category base-learners with sequential quadratic optimization ensemble (Li, Z., Goebel, K., *et al*, 2019)	29.51
Multi-category base-learners with particle swarm optimization ensemble (Li, Z., Goebel, K., *et al*, 2019)	31.62

Appendix B: Figures

Figure 1: Simplified diagram of the aircraft engine simulated in CMAPSS.



(Saxena. A., *et al*, 2008)

Figure 2: Proposed physics-guided ensemble learning progonostic for turbo-engine.

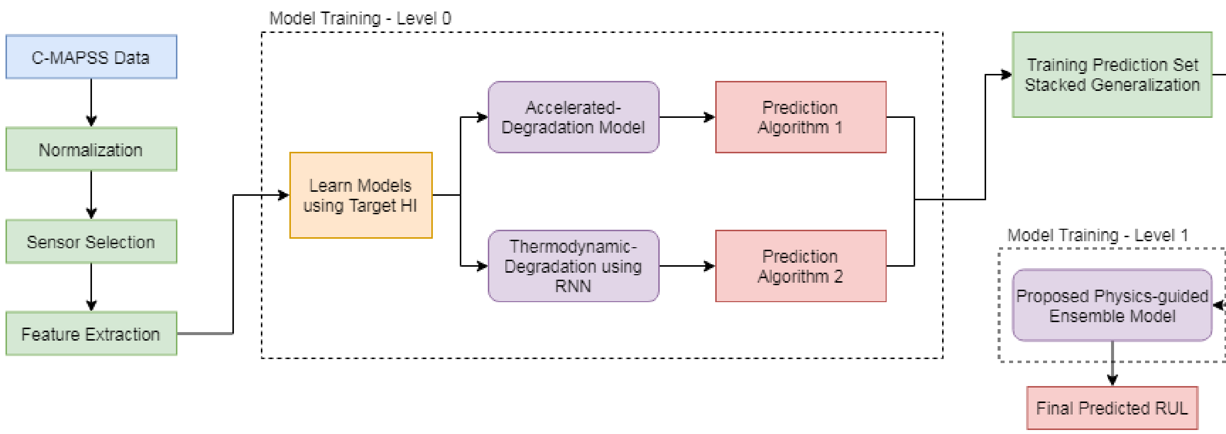


Figure 3: Raw Sensor data changing over time.

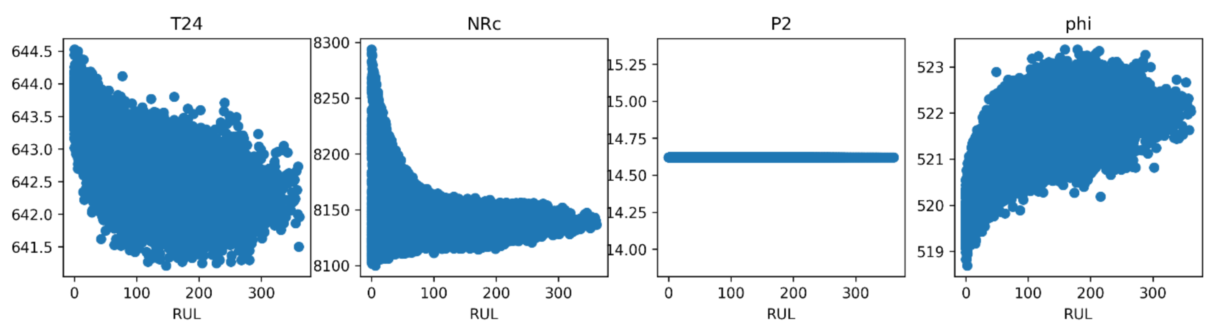


Figure 4: Illustration of the data during the entire life cycle of a single engine before and after normalization.

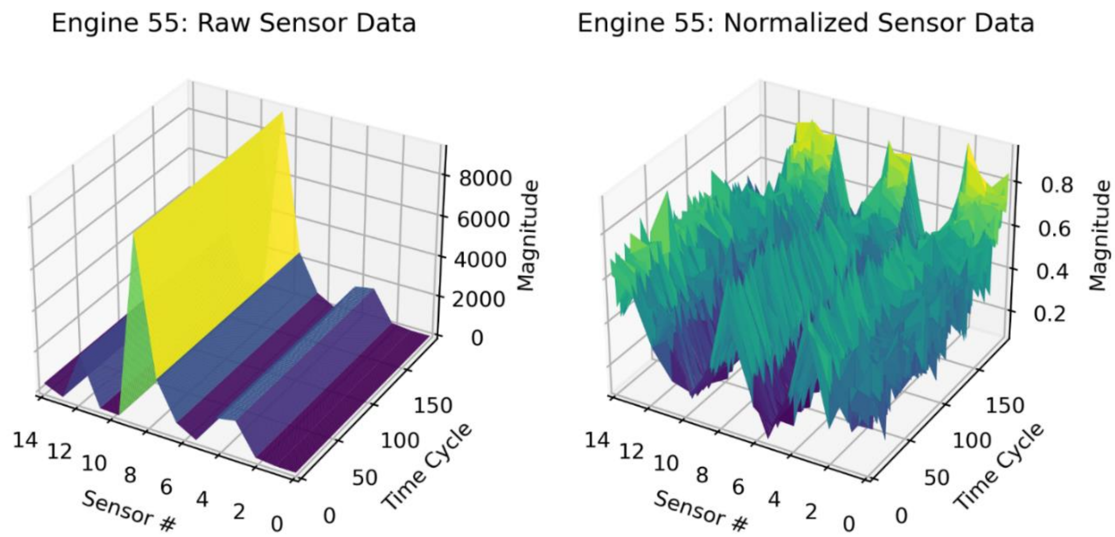


Figure 5: Variable Importance based on Random Forest model.

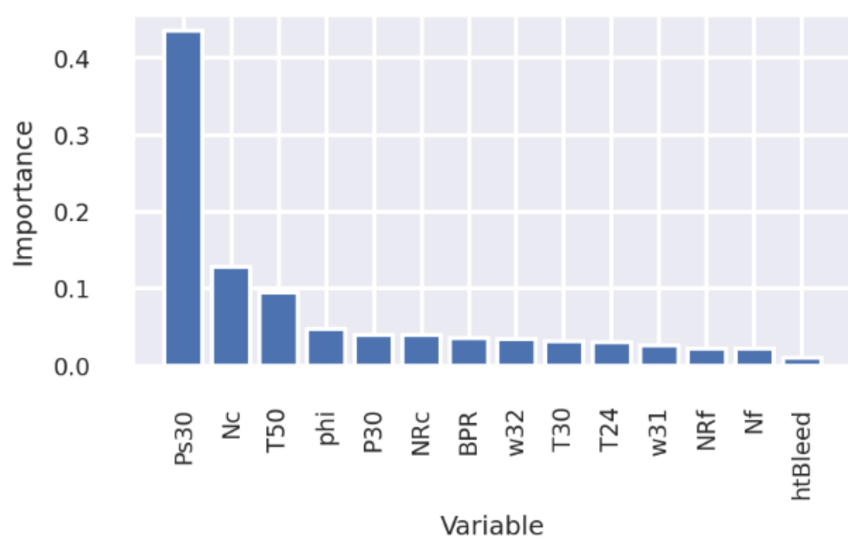


Figure 6: Correlation Matrix for Feature Selection.

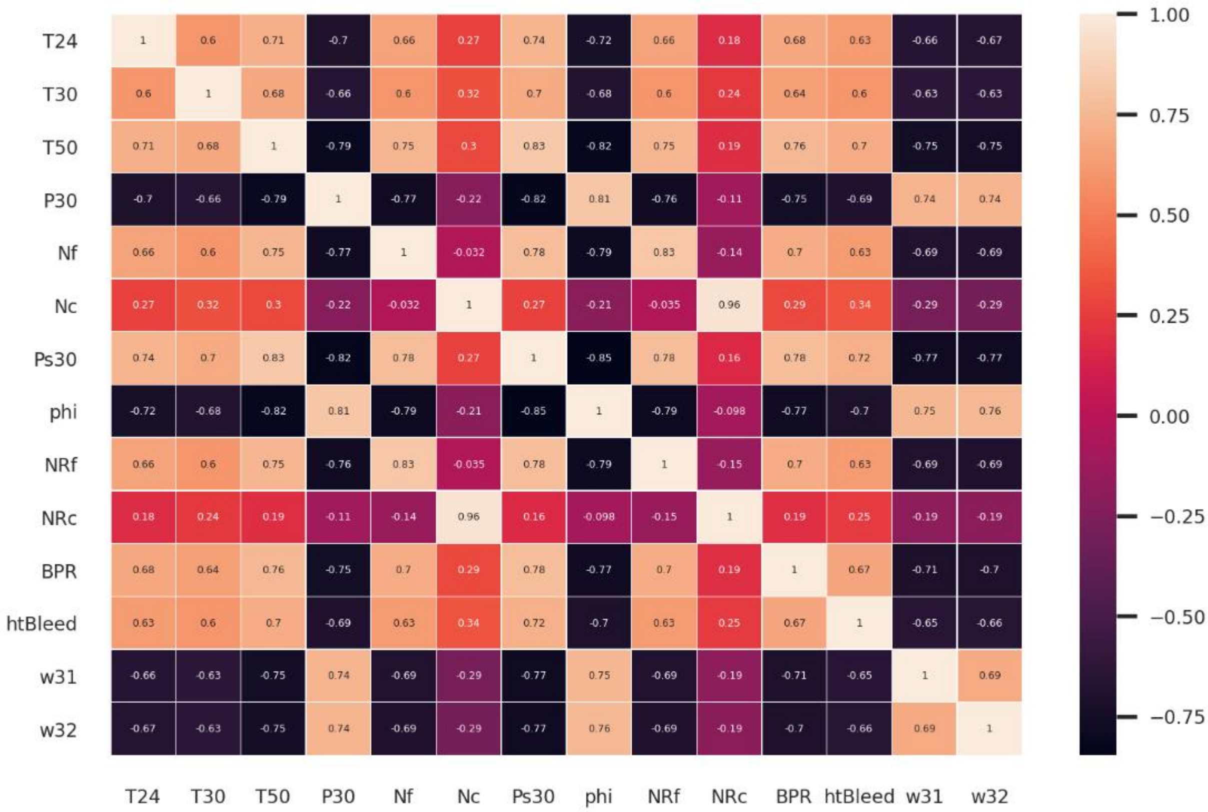


Figure 7: RUL estimates for test set FD001 given by the proposed ensemble model.