Graph Neural Networks for Leveraging Industrial Equipment Structure: An application to Remaining Useful Life Estimation

Jyoti Narwariya, Pankaj Malhotra, Vishnu TV, Lovekesh Vig, Gautam Shroff
jyoti.narwariya@tcs.com, malhotra.pankaj@tcs.com, vishnu.tv@tcs.com, lovekesh.vig@tcs.com, gautam.shroff@tcs.com
TCS Research, New Delhi, India

Abstract

Automated equipment health monitoring from streaming multi-sensor time series data can be used to enable condition-based maintenance, avoid sudden catastrophic failures, and ensure high operational availability. We note that most complex machinery has a well-documented and readily accessible underlying structure capturing the inter-dependencies between sub-systems or modules. Deep learning models such as those based on recurrent neural networks (RNNs) or convolutional neural networks (CNNs) fail to explicitly leverage this potentially rich source of domain-knowledge into the learning procedure. In this work, we propose to capture the structure of a complex equipment in the form of a graph, and use graph neural networks (GNNs) to model multi-sensor time series data. Using remaining useful life estimation as an application task, we evaluate the advantage of incorporating the graph structure via GNNs on the publicly available turbofan engine benchmark dataset. We observe that the proposed GNN-based RUL estimation model compares favorably to several strong baselines from literature such as those based on RNNs and CNNs. Additionally, we observe that the learned network is able to focus on the module (node) with impending failure through a simple attention mechanism, potentially paving the way for actionable diagnosis.

1 Introduction

Complex industrial equipment such as engines, turbines, aircrafts, etc., are typically instrumented with a large number of sensors that result in multivariate time series data. Most deep learning approaches model such multivariate time series data using variants of recurrent neural networks (RNNs) and convolutional neural networks (CNNs), e.g. (Heimes 2008; Malhotra et al. 2015; Babu, Zhao, and Li 2016; Zhang et al. 2016; Li, Ding, and Sun 2018). These approaches often follow an "end-to-end" design philosophy which emphasizes minimal a priori assumptions about the system (LeCun, Bengio, and Hinton 2015), and therefore, ignore or fail to leverage explicit structures. However, in most industrial setups, a complex equipment has a well-defined and documented structure: it consists of multiple interconnected modules, with the dynamics of one module affecting the dynamics of other modules. Fig. 1 shows an aircraft turbofan engine with several interconnected modules. Existing deep learning approaches fail to leverage the underlying structure of the complex equipment, and do not have the explicit capacity to reason about inter-component relations to make decisions over a structured representation of the sensor data.

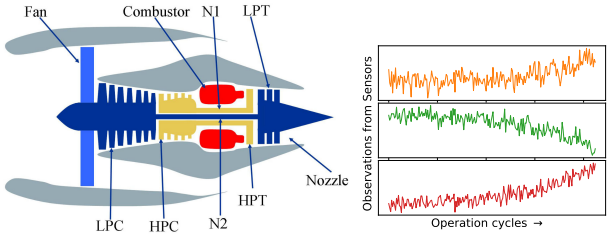


Figure 1: Left: Simplified structure of a turbofan engine depicting different modules / components [adapted from (Saxena et al. 2008)]. Right: Observations from sensors installed on an instance of the engine depicting typical degradation trends due to an impending failure.

Recently, a class of models for modeling and reasoning over graphs have been proposed. These include graph neural networks (GNNs) (Scarselli et al. 2008; Li et al. 2015), and their recent generalization in form of graph networks (Battaglia et al. 2018). This class of neural networks operate on graphs, and structure their computations accordingly. GNNs provide the desired *relational inductive bias* (Battaglia et al. 2018; Hamrick et al. 2018) to solve problems with underlying structure. For instance, NerveNet (Wang et al. 2018) uses Gated GNNs (GGNN) (Li et al. 2015) to learn structured policies by explicitly modeling the structure of the agent, and are better than the policies learned by models such as multi-layer perceptrons (MLPs) that simply concatenate all the observations from the environment.

In this work, we explore the applicability of GGNNs to explicitly model the underlying graph-structured mechanism of IoT-enabled complex equipment. We represent the structure of a particular complex equipment as a directed graph where each node corresponds to a subset of sensors (e.g. those from the same module), and an edge models the re-

lationship or dependency between two nodes or subsets of sensors (e.g. the dependence of a module on another module). Effectively, the multivariate time series of sensor data is then represented in the graph domain to learn a GGNN model. We use remaining useful life (RUL) estimation (Si et al. 2011) as a target application to validate our approach.

The key contributions of this work can be summarized as follows:

- We propose an approach to capture the knowledge of the structure of a complex equipment by using GGNNs. To the best of our knowledge, this is the first study to evaluate the effect of introducing such relational inductive bias into the deep learning approaches for equipment health monitoring.
- We show the advantage of informed modularized processing of the multi-sensor time series data by grouping the sensors into meaningful subsets guided by the underlying graph structure and modules, rather than the commonly used approach that concatenates the observations from all sensors into one multivariate time series.
- We provide insights into the working of GGNNs for RUL estimation: we observe that the modularized processing of multivariate time series using GGNNs followed by a simple attention mechanism for aggregating information across modules can potentially allow the network to focus more on the modules with impending failures.

2 Related Work

Recent works in (Sanchez-Gonzalez et al. 2018; Wang et al. 2018) implement an inductive bias for object- and relation-centric representations of complex dynamical systems such as a pendulum, cartpole, toy cheetah, etc. In this work, we draw inspiration from such approaches, and show that leveraging such inductive bias can improve performance in IIoT-enabled health monitoring applications, e.g. RUL estimation. To the best of our knowledge, this is the first attempt to model the time series of multivariate time series data for equipment health monitoring applications while leveraging the underlying structure and semantics of the equipment.

Another line of research focuses on incorporating the semantics of the problem into the structure of deep learning models by using ontologies. For instance, (Huang, Zanni-Merk, and Crémilleux 2019) propose using dense layers connected as per the ontology of the manufacturing line, followed by an RNN at the top to capture the temporal dependencies. Similarly, (Zhang et al. 2019) attempt to model an equipment as a graph of sensor nodes: they assume a fully-connected graph where nodes correspond to sensors, and edges capture the dependencies across sensors. However, they do not explicitly model the dependence between various modules of an equipment. Our work can be seen as a generalization of these approaches as it uses the structure of equipment to guide data processing.

Several variants of deep neural networks including CNNs and RNNs have been proposed for equipment health monitoring and RUL estimation, e.g. (Heimes 2008; Malhotra et al. 2015; Babu, Zhao, and Li 2016; Li, Ding, and Sun 2018; Gugulothu et al. 2017). However, most of these approaches consider a flat concatenation of readings or observations from all the sensors as a multivari ate input to the neural network, and ignore the structure of the underlying system or mechanism from which the data is generated. In this work, we show that grouping the sensors into meaningful inter-dependent subsets and processing them separately before the final concatenation step yields superior performance.

3 Problem Setup

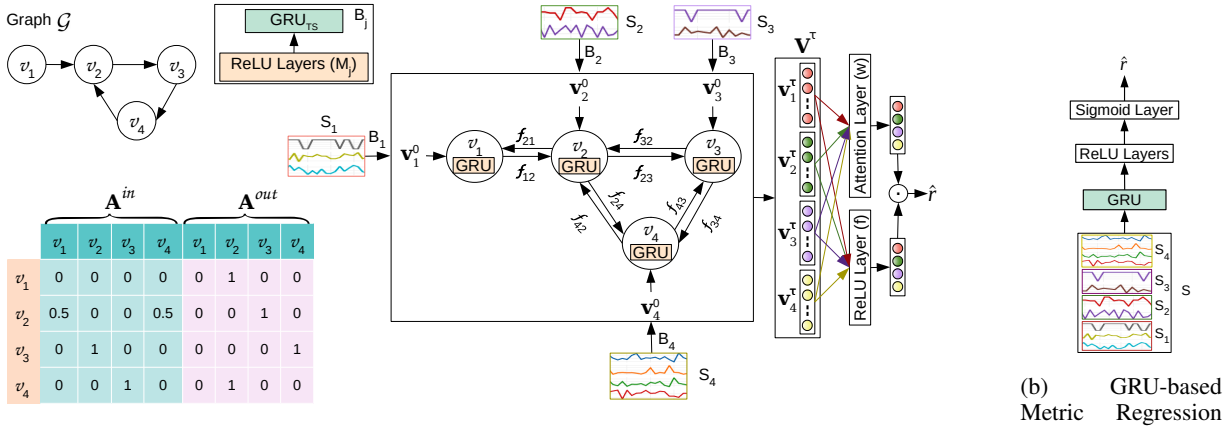
We consider the scenario where a complex equipment consists of multiple modules (sub-systems) connected to each other in a known fashion. Let $\mathcal{S}$ denote the set of all the sensors installed to monitor various parameters across various modules of the equipment. The dynamical behavior of any module is observed via the multivariate time series corresponding to a fixed and known subset of sensors (parameters) associated with that module. We represent the equipment as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $v_j \in \mathcal{V}$ (for $j = 1 \dots |\mathcal{V}|$) is a node in the graph that corresponds to a subset of sensors $\mathcal{S}_j \subset \mathcal{S}$ associated with the module indexed by j, $e_{jk} = (v_j, v_k) \in \mathcal{E}$ is a directed edge from node v_j to v_k that models the influence of $\mathcal{S}_k$ on $\mathcal{S}_j$. Note that, in general, $\mathcal{S}_j \cap \mathcal{S}_k \neq \emptyset$, such that a sensor can be associated with more than one node: for example, a sensor measuring the ambient temperature can be associated with all nodes.

We consider the supervised learning setting: we are given a fixed graph structure $\mathcal{G}$ and a learning set $\mathcal{D} = \{\mathbf{x}^i_1 \dots \mathbf{x}^i_{|\mathcal{V}|}, r^i\}_{i=1}^n$ of n time series, where $r^i \in \mathbb{R}$ is the target value, and $\mathbf{x}^i_1 \dots \mathbf{x}^i_{|\mathcal{V}|}$ denotes the $|\mathcal{V}|$ multivariate time series associated with the $|\mathcal{V}|$ nodes of $\mathcal{G}$. Here, $\mathbf{x}^i_j = \mathbf{x}^i_{j,1} \dots \mathbf{x}^i_{j,T}$ denotes the p_j-dimensional multivariate time series corresponding to node v_j for time $t = 1 \dots T$, where $\mathbf{x}^i_{j,t} \in \mathbb{R}^{p_j}$ and $p_j = |\mathcal{S}_j|$ denote the number of sensors in $\mathcal{S}_j$.

For the RUL estimation task, the n time series are collected from one or more instances (installations) of an equipment with structure $\mathcal{G}$, and the target variable $r^i \in \mathbb{R}$ corresponds to the RUL value at time T^i, e.g. in terms of remaining cycles of operation or remaining operational hours. RUL estimation is then a metric regression task where the goal is to map $\mathbf{x}^i_1 \dots \mathbf{x}^i_{|\mathcal{V}|}$ to r^i. Let F^i denote the total operational life of an instance i till the failure point, s.t. at any time $T^i \leq F^i$, the target RUL is given by $r^i = F^i - T^i$. Furthermore, as in a typical practical setting, we assume that all instances of the equipment have the same underlying graph structure $\mathcal{G}$, i.e. the different modules of the equipment are connected to each other in same fashion.

4 Approach

As illustrated in Fig. 2a, each multivariate time series $\mathbf{x}_j$ is processed by a neural network B_j (for $j = 1, \dots, |\mathcal{V}|$) to obtain a fixed-dimensional initial node representation vector $\mathbf{v}^0_j \in \mathbb{R}^d$. This initial representation $\mathbf{v}^0_j$ is then updated using the representations of neighboring nodes defined by $\mathcal{G}$ using a message passing mechanism to obtain $\mathbf{v}^\tau_j$. Finally, an attention mechanism is used to combine the final node representations to obtain an RUL estimate $\hat{r}$ for the equipment.



(a) GGNN-based Metric Regression (GNMR)

(b) GRU-based Metric Regression (GRU-MR)

Figure 2: Illustrative process flow for the proposed GGNN-based Metric Regression (GNMR) and the traditional GRU-based Metric Regression. While GRU-MR ingests observations from all sensors at once, GNMR processes observations grouped based on the equipment structure. f_{ij} refers to one or more feedforward layers with leaky ReLUs that capture the influence of node j on node i. $\mathbf{v}_j^0$ denotes the initial representation of node j and $\mathbf{v}_j^\tau$ denotes the final representation of node j after τ propagation steps. (Best viewed in the electronic version by zooming-in.)

We refer to this approach as **GNMR** (Gated **G**raph Neural **N**etworks for **M**etric **R**egression). The entire computation flow of GNMR is differentiable end-to-end and the associated parameters are learned via stochastic gradient descent. Next, we describe these steps in more detail.

Learning Node Representations from Time Series

The p_j-dimensional time series $\mathbf{x}_j$ at node v_j is processed by B_j to obtain $\mathbf{v}_j^0$. We consider a gated recurrent units (GRU)-based RNN (Cho et al. 2014) for processing this time series. In general, p_j is different across the nodes implying that we need to learn and maintain $|\mathcal{V}|$ GRU networks. This can pose scalability issues for graphs with large number of nodes. It is, therefore, desirable to use a common GRU, which we refer to as GRU_{TS}, to process the multivariate time series from all the nodes so as to keep the number of trainable parameters of the network within manageable limits. To this end, any point $\mathbf{x}_{j,t} \in \mathbb{R}^{p_j}$ $(t = 1 \dots T)$ at node v_j is first processed via a node-specific feedforward network M_j to obtain a fixed d-dimensional vector $\tilde{\mathbf{x}}_{j,t} \in \mathbb{R}^d$. Note that d is same across nodes allowing us to use the common GRU_{TS} for further processing of the resulting time series $\tilde{\mathbf{x}}_j$ to obtain the initial representation $\mathbf{v}_j^0$. Effectively, B_j consisting of M_j and GRU_{TS} maps the input time series $\mathbf{x}_j$ to $\mathbf{v}_j^0$.

Message Passing Across Nodes

While the node-level representation $\mathbf{v}_j^0$ obtained from B_j can capture the dependencies across sensors in $\mathcal{S}_j$, it ignores the dependencies across nodes. It is desirable to leverage the representations of neighboring nodes to capture the dependencies between nodes, and then aggregate them to obtain a representation for the overall dynamics of the system. To achieve this, the representations for each node are iteratively updated by a GGNN using the representations of the neighboring nodes as described next.

Consider two normalized adjacency matrices $\mathbf{A}^{in} \in \mathbb{R}^{|\mathcal{V}|\times\mathcal{V}|}$ and $\mathbf{A}^{out} \in \mathbb{R}^{|\mathcal{V}|\times\mathcal{V}|}$ corresponding to the incoming and outgoing edges in graph $\mathcal{G}$, as illustrated in Fig. 2a. Let $\mathbf{v}_j^m \in \mathbb{R}^d$ correspond to the jth row of matrix $\mathbf{V}^m \in \mathbb{R}^{|\mathcal{V}|\times d}$, and denote the representation (or embedding) for node v_j after m message propagation steps. GGNN takes $\mathbf{A}^{in}$, $\mathbf{A}^{out}$, and the initial node representations $\mathbf{V}^0$ as input, and returns an updated set of representations $\mathbf{V}^\tau$ after τ iterations of message propagation across nodes in the graph s.t. $[\mathbf{v}_1^\tau, \mathbf{v}_2^\tau, \dots, \mathbf{v}_{|\mathcal{V}|}^\tau] = G(\mathbf{A}^{in}, \mathbf{A}^{out}, \mathbf{V}^0; \boldsymbol{\theta}_g)$, where $\boldsymbol{\theta}_g$ represents the parameters of the GGNN function G. For any node v_j in the graph and message propagation step m, the previous representation of the node $\mathbf{v}_j^{m-1}$, and the aggregated representation $\mathbf{a}_j^m$ of its neighboring nodes (as obtained via Eqs. 1-3 below) are used to iteratively update the representation of the node τ times.

More specifically, the representation of node v_j in the message propagation step m $(= 1 \dots \tau)$ is updated as follows:

$$\mathbf{p}_{ij}^m = f_{ij}(\mathbf{v}_i^{m-1}; \boldsymbol{\theta}_{ij}),\ \mathbf{p}_{jk}^m = f_{jk}(\mathbf{v}_k^{m-1}; \boldsymbol{\theta}_{jk}), \quad (1)$$

$$\mathbf{P}_{1j}^m = [\mathbf{p}_{1j}^m \cdots \mathbf{p}_{|\mathcal{V}|j}^m]^\top,\ \mathbf{P}_{2j}^m = [\mathbf{p}_{j1}^m \cdots \mathbf{p}_{j|\mathcal{V}|}^m]^\top, \quad (2)$$

$$\mathbf{a}_j^m = [\mathbf{A}_{j:}^{in}\mathbf{P}_{1j}^m, \mathbf{A}_{j:}^{out}\mathbf{P}_{2j}^m]^\top, \quad (3)$$

$$\mathbf{z}_j^m = \sigma(\mathbf{W}_z\mathbf{a}_j^m + \mathbf{U}_z\mathbf{v}_j^{m-1}), \quad (4)$$

$$\mathbf{r}_j^m = \sigma(\mathbf{W}_r\mathbf{a}_j^m + \mathbf{U}_r\mathbf{v}_j^{m-1}), \quad (5)$$

$$\hat{\mathbf{v}}_j^m = \tanh(\mathbf{W}_o\mathbf{a}_j^m + \mathbf{U}_o(\mathbf{r}_j^m \odot \mathbf{v}_j^{m-1})), \quad (6)$$

$$\mathbf{v}_j^m = (1 - \mathbf{z}_j^m) \odot \mathbf{v}_j^{m-1} + \mathbf{z}_j^m \odot \hat{\mathbf{v}}_j^m, \quad (7)$$

where $i, k = 1, \dots, |\mathcal{V}|$, $\mathbf{A}_{j:}^{in}$ and $\mathbf{A}_{j:}^{out}$ denote the j-th row of $\mathbf{A}^{in}$ and $\mathbf{A}^{out}$, respectively. Here, $\mathbf{A}^{in}$ and $\mathbf{A}^{out}$ allow

to capture the information from upstream and downstream nodes in the system, respectively. f_{ij} denotes feedforward ReLU layer(s) with parameters $\boldsymbol{\theta}_{ij}$ that computes the contribution (message) from v_i to v_j if there is an incoming edge from v_i to v_j, i.e. when $e_{ij} \in \mathcal{E}$. Then, $\mathbf{p}_{ij}^m \in \mathbb{R}^d$ denotes the message from v_i to v_j corresponding to edge e_{ij}. Similarly, f_{jk} computes the message from v_k to v_j if there is an outgoing edge from v_j to v_k, i.e. when $e_{jk} \in \mathcal{E}$. $\mathbf{P}_{1j}^m$ and $\mathbf{P}_{2j}^m \in \mathbb{R}^{|\mathcal{V}| \times d}$ denote the matrices that contain the information from the incoming and outgoing edges with v_j as starting and ending node, respectively. For $e_{ij} \notin \mathcal{E}$, f_{ij} simply returns $\mathbf{0} \in \mathbb{R}^d$. The trainable parameters $\boldsymbol{\theta}_{ij}$, $\boldsymbol{\theta}_{ji}$, $\mathbf{W}_{(.)}$ and $\mathbf{U}_{(.)}$ of appropriate dimensions constitute $\boldsymbol{\theta}_g$, $\sigma(.)$ is the sigmoid function, and $\odot$ is the element-wise multiplication operator. Eqs. 4-7 are the computations equivalent to a gated recurrent unit (GRU) network.

Note that for large graphs with many nodes and edges, the number of functions f_{ij} and their associated parameters $\boldsymbol{\theta}_{ij}$ can be large. However, in many practical applications such as a power plant or a water treatment plant (Goh et al. 2016), nodes typically have an associated type with more than one node belonging to the same type, e.g. multiple water tanks in a water treatment plant. Under such scenarios, it may be suitable to tie the parameters of the edge functions for a given pair of node types.

Attention Mechanism to Aggregate Node-level Representations

The final representations $\mathbf{v}_1^\tau, \ldots, \mathbf{v}_{|\mathcal{V}|}^\tau$ can be aggregated to get a graph-level output (RUL estimate in our case) by using an attention mechanism, e.g. as used in (Li et al. 2015). In this work, we consider a simple variant of this attention mechanism: For each node, we use the concatenated vector $\tilde{\mathbf{v}}_j = [\mathbf{v}_j^0, \mathbf{v}_j^\tau, T, node\_type_j]$ as inputs to two parallel feedforward layers f_1 and f_2 to obtain $f_1(\tilde{\mathbf{v}}_j^\tau) \in \mathbb{R}$ and $\hat{r}_j = f_2(\tilde{\mathbf{v}}_j^\tau) \in \mathbb{R}$. Here, $node\_type_j$ is a one-hot vector of length $|\mathcal{V}|$, and is set to 1 for jth position, and 0 otherwise. Also, T is used as an additional input for the RUL estimation task as the total life passed can be a useful feature to estimate the wear-and-tear of the system. We apply softmax over the values from f_1 to obtain attention weight $w_j = \frac{\exp(f_1(\tilde{\mathbf{v}}_j^\tau)}{\sum_i \exp(f_1(\tilde{\mathbf{v}}_i^\tau))}$ for node v_j. The final RUL estimate is then given by

$$\hat{r} = \sum_{j=1}^{|\mathcal{V}|} w_j \hat{r}_j. \tag{8}$$

This can be interpreted as assigning a weightage $0 \leq w_j \leq 1$ to the node v_j while $\hat{r}_j = f_2(\tilde{\mathbf{v}}_j^\tau)$ denotes the RUL estimate as per node v_j.

Given the training set $\mathcal{D}$, the loss function used for training is then given by $\mathcal{L}_\mathcal{D} = \frac{1}{n}\sum_i^n (r^i - \hat{r}^i)^2$, with learning parameters being $\boldsymbol{\theta}_g$, parameters of GRU_{TS}, and the parameters of the feedforward layers (M_is, f_1, and f_2) with leaky ReLU units.

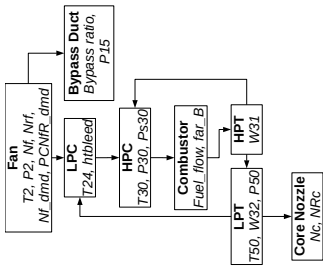


Figure 3: Original graph structure used for the experiments. Each node corresponds to a module (in bold) with an associated subset of sensors (in italics). Refer Table 2 in (Saxena et al. 2008) for details of sensor names.

5 Experimental Evaluation

We investigate if **GNMR** is able to leverage the graph structure to improve upon other strong baselines from literature that consider a simple concatenation of observations as a multivariate input: CNNs (Babu, Zhao, and Li 2016; Li, Ding, and Sun 2018), LSTMs (Zheng et al. 2017), and Deep Belief Networks (Zhang et al. 2016). We additionally implement **GRU-MR**: a GRU-based RUL estimation model on lines of Metric Regression approach proposed in (Zheng et al. 2017) as depicted in Fig. 2b. We ensure comparable hyperparameter settings as well as the same train, validation and test splits for GRU-MR and the proposed GNMR, as detailed later. We further study the sensitivity of the approach to the knowledge of the graph structure by synthetically combining or segregating nodes (modules) in the original graph structure. To study other ways of combining time series of sensors, we also consider reducing the number of input dimensions for GRU via principal components analysis within the GRU-MR framework, and refer to it as **PCA-GRU-MR**. We report results for 5 PCA components (capturing 85% variance in the data), i.e. 5-dimensional input to GRU-MR in Table 1.

Datasets: We use the four publicly available aircraft turbofan engine benchmark datasets[1] FD001-FD004, as introduced in (Saxena et al. 2008). We use the equipment structure information as depicted in Figs. 1 and 3 (refer (Saxena et al. 2008) for details). Each dataset contains a pre-defined train-test split. We further use a random 80-20 split of the original train split to obtain a train and a validation set. The hold-out validation set is used for hyperparameter tuning.

Performance Metrics: We use the standard metrics RMSE and Timeliness Score (S) as introduced in (Saxena et al. 2008). Let $e^i = \hat{r}^i - r^i$ denote the error in RUL estimate for ith test instance, then $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^n (e^i)^2}$ and $S = \sum_{i=1}^n (\exp(\gamma.|e^i|) - 1)$, n is the number of test instances, $\gamma = 1/u_1$ if $e^i < 0$, else $\gamma = 1/u_2$. Usually, $u_1 > u_2$ such that late predictions are penalized more compared to early predictions. We consider $u_1 = 13$ and $u_2 = 10$ as used in all the baselines. Lower values of RMSE

[1] https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#turbofan

Table 1: Performance comparison in terms of RMSE and S. Best performance is shown in bold and second-best is underlined.

Dataset →	FD001		FD002		FD003		FD004		Average Rank	
Method ↓	RMSE	S	RMSE	S	RMSE	S	RMSE	S	RMSE	S
CNN-MR (Babu, Zhao, and Li 2016)	18.45	1,287	30.29	13,570	19.82	1596	29.16	7,886	8	7.75
LSTM-MR (Zheng et al. 2017)	16.14	338	24.49	4,450	16.18	852	28.17	5,550	6.5	5.25
MODBNE (Zhang et al. 2016)	15.04	334	25.05	5,585	**12.51**	422	28.66	6,558	4.75	5.25
CNN + FNN (Li, Ding, and Sun 2018)	12.61	274	22.36	10,412	12.64	**284**	23.31	12,466	3.25	4.5
GRU-MR (ours)	15.36	481	22.43	3,391	12.52	339	22.96	2,964	3.75	3.75
PCA-GRU-MR (ours)	15.60	469	22.92	3,916	13.37	860	22.41	**2,637**	5	4.5
GNMR (proposed, $\tau = 0$)	12.73	302	21.38	**3,148**	13.06	366	21.81	3,414	2.75	2.75
GNMR (proposed)	**12.14**	**212**	**20.85**	3,196	13.23	370	**21.34**	2,795	**2**	**2.25**

Table 2: Effect of varying the graph structure.

Dataset →		FD001		FD002		FD003		FD004		Average Rank	
Method ↓	$\lvert\mathcal{V}\rvert$	RMSE	S	RMSE	S	RMSE	S	RMSE	S	RMSE	S
Original Graph	8	**12.14**	**212**	**20.85**	**3,196**	**13.23**	**370**	21.34	2,795	**1.25**	**1.5**
One node for all sensors	1	13.20	321	22.36	4,072	13.65	378	22.04	2,747	4.25	3
Reduced Nodes	4	12.41	299	22.85	4,798	15.26	981	21.87	3,034	4.25	4.5
Increased Nodes	13	12.15	214	22.20	4,512	13.63	623	**20.95**	3,813	2.25	3.75
One node per sensor	21	13.86	293	21.99	3,317	13.39	431	21.48	**2,562**	3	2.25

and S indicate better performance.

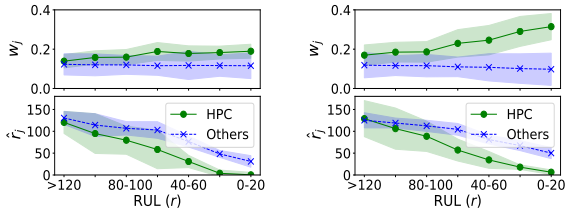
Hyperparameters: We use batch size of 32, and a dropout rate of 0.2 for all feedforward (leaky ReLU) layers. We use 2 leaky ReLU layers for each M_j and f_{ij} in GNMR as well as for the pre-final leaky ReLU layers of GRU-MR (refer Fig. 2). We use Adam optimizer with initial learning rate of 0.001 which is reduced every 10 epochs by a factor of $1/\sqrt{2}$. The number of hidden units, same as d, for all feedforward and recurrent layers is chosen from {30,60}. We use message propagation steps $\tau = \{0, 2, 4\}$ for GGNN. The number of hidden layers for GRU_{TS} in GNMR and for GRU-MR is chosen from {2,3,4}. All hyperparameters are selected via grid search based on validation RMSE.

Results and Observations

(1) *Comparison with baselines*: From Table 1, we observe that GGNN performs better than GRU-MR, PCA-GRU-MR as well as other baselines across most datasets. GNMR has the highest average rank of 2.0 based on both RMSE and S. Furthermore, the special case of GNMR with no message propagation across nodes, i.e. $\tau = 0$, also compares favorably to GRU-MR and other baselines. These results suggest that meaningful grouping of the sensors into nodes based on the knowledge of the modular graph structure of the equipment is advantageous over methods that consider a concatenated vector of all the sensors as one input. Also, allowing for message propagation across nodes gives further improvement in results indicating the advantage of modeling the dependencies between modules.

(2) *Effect of varying the graph structure*: We consider two scenarios to study the sensitivity of results to the exact knowledge of graph structure: for the *Increased Nodes* scenario, any original node v_j in $\mathcal{G}$ with $|\mathcal{S}_j| > 1$ is split into two nodes, say v_j^1 and v_j^2, by randomly distributing the sensors in $\mathcal{S}_j$ to the two nodes such that each new node gets half the sensors. Furthermore, the nodes v_j^1 and v_j^2 thus created are connected to each other. Also, if $e_{jk} \in \mathcal{E}$, then both v_j^1 and v_j^2 are additionally connected to the new nodes v_k^1 and v_k^2 obtained from v_k. For the *Reduced Nodes* scenario, when combining two neighboring nodes v_j and v_k into a new node, say v_{jk}, we have $\mathcal{S}_{jk} = \mathcal{S}_j \cup \mathcal{S}_k$, and an edge exists between any two new nodes if there was an edge between the original nodes from which the new nodes were created. We also consider the limiting cases of *One node per sensor* and *One node for all sensors* for the *Increased Nodes* and *Reduced Nodes* scenarios, respectively.

From Table 2, we observe that the performance degrades as we reduce the nodes in the graph, with *Reduced Nodes* and *One node for all sensors* being the worst models. On the other hand, the graph with *Increased Nodes* performs the same as the *Original Graph* for FD001 while being better on FD004. Nevertheless, the *Original Graph* still gives best performance on average across datasets. *Increased Nodes* and *One node per sensor* have the second-best performance. However, it is to be noted that the models with increased nodes have a much larger number of MLPs for message propagation (refer Eqs. 1-2) due to increased number of edges, making it computationally more expensive when compared to the original graph.



(a) FD001 (b) FD002

Figure 4: Average attention weight and RUL estimates w.r.t ground truth RUL for the HPC node with impending failure versus other nodes.

(3) *Preliminary Analysis of Attention Mechanism (refer Eq. 8)*: For FD001 and FD002 datasets, the faults in all instances

are known to originate in the HPC module[2] (refer Fig. 3). It is therefore expected that as the degradation increases, the behavior of sensors associated with the HPC module will depict signatures for detecting the impending failure, and in turn, estimating the RUL. We observe that GNMR implicitly tends to capture and leverage this behavior in some cases: we analyze the attention weights w_js and the corresponding contribution $\hat{r}_j$ to the RUL estimate from the HPC node versus the remaining nodes. As shown in Fig. 4(b), we observe that for FD002 dataset, the attention for the faulty node (HPC module) increases as the target RUL decreases (i.e. as the engines approach failure), while the attention tends to decrease for the remaining nodes. However, this trend is not observed in FD001 Fig. 4(a). In future, it will be interesting to see if this can be explicitly ensured: while other modules may provide additional information to track the health degradation of a particular module, it may be useful to bias the attention to the module-of-interest.

6 Conclusion and Future Work

We have proposed an approach to incorporate the readily available information about the modularized structure of complex equipment into deep learning models via gated graph neural networks (GNNs). To the best of our knowledge, our work provides a first set of results for leveraging GNNs in the increasingly important area of automated equipment health monitoring. We analyze the heavily-benchmarked aircraft turbofan engine dataset through the lens of structure-aware deep learning, potentially bridging the gap between deep learning and the domain-knowledge aware approaches. We hope that this work inspires future research in leveraging equipment structure to model the behavior of complex systems (such as power plants) with interesting applications like optimization and anomaly detection. While the graph structure is readily available in most practical applications as part of domain knowledge, it may not be optimal in terms of reflecting the dependencies between sensors at a node or between sensors across nodes. It will be interesting to explore if the optimal graph structure can itself be learned starting from the domain knowledge-based initial graph structure.

References

Babu, G. S.; Zhao, P.; and Li, X.-L. 2016. Deep convolutional neural network based regression approach for estimation of remaining useful life. In *International conference on database systems for advanced applications*, 214–228. Springer.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Goh, J.; Adepu, S.; Junejo, K. N.; and Mathur, A. 2016. A dataset to support research in the design of secure water treatment systems. In *International Conference on Critical Information Infrastructures Security*, 88–99. Springer.

Gugulothu, N.; TV, V.; Malhotra, P.; Vig, L.; Agarwal, P.; and Shroff, G. 2017. Predicting remaining useful life using time series embeddings based on recurrent neural networks. *arXiv preprint arXiv:1709.01073*.

Hamrick, J. B.; Allen, K. R.; Bapst, V.; Zhu, T.; McKee, K. R.; Tenenbaum, J. B.; and Battaglia, P. W. 2018. Relational inductive bias for physical construction in humans and machines. *arXiv preprint arXiv:1806.01203*.

Heimes, F. O. 2008. Recurrent neural networks for remaining useful life estimation. In *Prognostics and Health Management, 2008. PHM 2008.*, 1–6. IEEE.

Huang, X.; Zanni-Merk, C.; and Crémilleux, B. 2019. Enhancing deep learning with semantics: an application to manufacturing time series analysis. *Procedia Computer Science* 159:437–446.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553):436–444.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Li, X.; Ding, Q.; and Sun, J.-Q. 2018. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety* 172:1–11.

Malhotra, P.; Vig, L.; Shroff, G.; and Agarwal, P. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. In *ESANN, 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 89–94.

Sanchez-Gonzalez, A.; Heess, N.; Springenberg, J. T.; Merel, J.; Riedmiller, M.; Hadsell, R.; and Battaglia, P. 2018. Graph networks as learnable physics engines for inference and control. *arXiv preprint arXiv:1806.01242*.

Saxena, A.; Goebel, K.; Simon, D.; and Eklund, N. 2008. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management*, 1–9. IEEE.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1):61–80.

Si, X.-S.; Wang, W.; Hu, C.-H.; and Zhou, D.-H. 2011. Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research* 213(1):1–14.

Wang, T.; Liao, R.; Ba, J.; and Fidler, S. 2018. Nervenet: Learning structured policy with graph neural networks.

Zhang, C.; Lim, P.; Qin, A. K.; and Tan, K. C. 2016. Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. *IEEE transactions on neural networks and learning systems* 28(10):2306–2318.

[2]For FD003 and FD004, the ground truth information about the faulty node is not available, hence not analyzed.

Zhang, W.; Zhang, Y.; Xu, L.; Zhou, J.; Liu, Y.; Gu, M.; Liu, X.; and Yang, S. 2019. Modeling iot equipment with graph neural networks. *IEEE Access* 7:32754–32764.

Zheng, S.; Ristovski, K.; Farahat, A.; and Gupta, C. 2017. Long short-term memory network for remaining useful life estimation. In *Prognostics and Health Management (ICPHM), 2017 IEEE International Conference on*, 88–95. IEEE.

Datasets and Pre-processing Details

The datasets contain time series of readings for 24 sensors (21 sensors and 3 operating condition variables), such that each cycle in the life of an engine provides a 24-dimensional vector. The sensor readings in the training set are available from beginning of usage of the engine till the end of life, while those in the test split are clipped at a random time prior to the failure such that the instances are operational at the last available cycle, and the goal is to estimate the RUL for these test instances. Refer Table 3 for details of the datasets.

Table 3: Basic statistics of the four datasets used.

Dataset →	**FD001**	**FD002**	**FD003**	**FD004**
Instances (training set)	80	208	80	199
Instances (validation set)	20	52	20	50
Instances (test set)	100	259	100	248
Operating conditions	1	6	1	6
Fault Modes	1	1	2	2

As commonly used in RUL estimation approaches (Zheng et al. 2017; Babu, Zhao, and Li 2016; Li, Ding, and Sun 2018), we also consider an upper bound r_u on the possible values of target RUL during training as, in practice, it is not possible to predict too far ahead in future. So if $r > r_u$, we clip the value of r to $r_u = 130$. The targets are then normalized to $\frac{r}{r_u}$, such that the targets in training are in the range 0-1. We use min-max normalization technique to normalize input time series sensor wise using the minimum and maximum value of each sensor (in the range of -1 to 1) from the training set. The time series for each engine instance is then divided into windows of length $T = 100$ with window-shift of 5, refer Table 4 for details of the number of resulting time series after windowing. We use suitable pre-padding with mean value of sensor readings to ensure same length for each time series for both GRU-MR and GGNN.

Table 4: Number of time series windows.

Dataset →	**FD001**	**FD002**	**FD003**	**FD004**
Training	1,875	4,688	2,129	5,383
Validation	411	1,287	533	1,451
Test	100	259	100	248