A semi-supervised RUL prediction with likelihood-based pseudo labeling for suspension histories

Ryosuke Takayama
Department of Aeronautics and Astronautics
University of Tokyo
Tokyo, Japan
ryo-tkym818@g.ecc.u-tokyo.ac.jp

Masanao Natsumeda
Digital Technology Development Laboratories
NEC Corporation
Kanagawa, Japan
mnatsumeda@nec.com

Takehisa Yairi
Research Center for Advanced Science and Technology
University of Tokyo
Tokyo, Japan
yairi@g.ecc.u-tokyo.ac.jp

2023 IEEE International Conference on Prognostics and Health Management (ICPHM) | 979-8-3503-4625-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICPHM57936.2023.10194226

***Abstract*—Accurate remaining useful life (RUL) prediction is an essential for efficient maintenance. In recent years, with the rapid development of industrial big data, many data-driven methods for RUL prediction have made significant progress, especially using deep learning. However, most of the proposed deep learning models only utilize labeled data and require a large amount of labeled data. In practice, the component of equipment is often replaced with a new one before it fails by preventive maintenance, resulting in a small number of failure histories and a large number of suspension histories. In other words, we have a small amount of labeled data and a large amount of unlabeled data. This paper proposes a new semi-supervised RUL prediction method using pseudo labels with flexibility in model architecture and low computational cost. For each suspension history, optimal pseudo labels are estimated using a likelihood-based method that takes into account important constraints, which enables more effective use of the information in both failure and suspension histories. The experiments on the C-MAPSS dataset validate the prediction accuracy of the proposed approach and provide several insights.**

***Index Terms*—remaining useful life prediction, deep learning, semi-supervised learning, prognostics and health management, C-MAPSS dataset**

I. Introduction

Maintenance of equipment is essential because various factors can cause equipment to deteriorate and not to work properly. Condition-Based Maintainance (CBM) is an efficient maintenance method that can prevent unexpected failures while reducing overall maintenance costs [1]. Accurate remaining useful life (RUL) prediction for equipment or its components is a vital technology for CBM and has been actively studied. In recent years, with the rapid development of industrial big data, many data-driven methods have been studied, and in particular, various approaches using deep learning have been investigated [2]. Typically, the models are trained with the condition monitoring data as input and the RUL as output, learning the relationship between them. For example, Cornelius et al. [3] proposed a model using 1D convolutional layers with temporal dilation factor to deal with long time series while keeping the number of required parameters low.

Although data-driven RUL prediction using deep learning has achieved some success, there are still some challenges. One of these challenges is that many of these RUL prediction methods are based on supervised learning and require a large amount of labeled data. There are two types of history data that can be obtained in the actual operation of the equipment that needs RUL prediction: failure history and suspension history. In failure history, the components in industrial systems are operated and their condition monitoring data is taken until they fail. Since the RUL at the failure time is considered to be 0, the RUL at each time point can be calculated from there, meaning that failure history is regarded as the labeled data. In suspension history, the components are replaced before a failure occurs (e.g. by preventive maintenance). Although their condition monitoring data is obtained up to that suspension time, the time when the failure should have occurred (failure time) is not known, meaning that we can not calculate the RUL at each time of suspension history and that it is regarded as unlabeled data. In general, as the unplanned downtime of the equipment due to failures is not tolerated, we commonly have a small number of failure histories and a large number of suspension histories. In other words, we have a small amount of labeled data and a large amount of unlabeled data, and in such cases, semi-supervised learning, which utilizes both labeled and unlabeled data, is effective. There have been some previous studies on semi-supervised RUL prediction. For instance, Yoon et al. [4] trained Variational Autoencoder (VAE) with both labeled and unlabeled data, and utilized the encoder of the VAE as the encoder and Recurrent Neural Networks (RNN) as the decoder. They predicted the RUL of suspended data based on the self-learning and retrained the model by using all histories with RUL labels. Hu et al. [5] constructed two regressors to predict the RUL for each other's unlabeled data and trained these regressors with the RUL labeled data. Although, in these studies, the use of suspension histories allows for more accurate RUL prediction, there are still some drawbacks. There are some constraints that should be satisfied as RULs of a suspension history. One is derived from the fact that one suspension history has one (unknown) failure time. Since the failure time is a reference determining RULs in the same history, they should be jointly estimated. Another is that the unknown failure time should be greater than the suspension time of that suspension history. These approaches cannot take these constraints into consideration. In [6], He et al. proposed a RUL prediction method that uses Generative Adversarial Network (GAN) to learn degradation manifolds from both failure and suspension histories while taking into account the above constraints. Although this approach showed excellent performance, there are considered to be some challenges specific to GAN, such as the complexity of the architecture and learning difficulties. Tian et al. [7] exploited suspension histories by assigning pseudo labels to them while ensuring that the predicted failure time of suspension histories is always greater than the suspension time. Although this proposed procedure is feasible for models of various architectures, it is computationally expensive because the model must be trained multiple times to predict the failure time for each suspension history.

In this paper, we propose a semi-supervised RUL prediction method using pseudo labels, which does not require multiple model trainings for pseudo labeling and can assign pseudo labels to all suspension histories at once. It is guaranteed that the pseudo labels assigned by this approach satisfy the above constraints. Moreover, by utilizing the information from not only the failure histories but also the suspension histories, it is possible to assign more correct pseudo labels. Since our proposal is involved in the training process including how to determine the pseudo labels, it is applicable to various model architectures as long as they satisfy several requirements.

The rest of the paper is organized as follows. In Section II, the proposed semi-supervised RUL prediction method is described. Section III provides the experiments using the NASA C-MAPSS dataset to validate the proposed method. Finally, the conclusions are given in Section IV.

II. Proposed Method

A. Overview

In this study, we propose a semi-supervised RUL prediction method utilizing suspension histories by assigning pseudo labels. There are two key ideas to improve the accuracy of the pseudo labels. One is to consider the constraints of suspension history by assigning a pseudo label sequence to each history, instead of assigning a pseudo label to each point in each history. As mentioned in Section I, all suspension histories

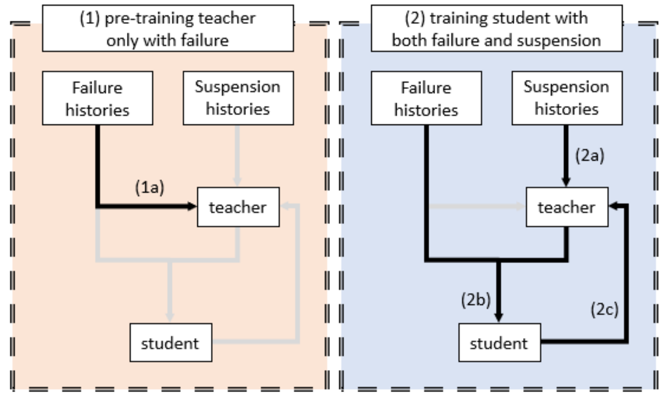

Fig. 1. Summary of training process.

and RUL histories assigned to them as pseudo labels have two constraints that should be satisfied: (i) each point in one suspension history has the same unknown failure time and the corresponding RUL is not independent, (ii) this unknown failure time should be greater than the suspension time of the suspension history. Therefore, for each suspension history, after assuming the predicted failure time that is greater than the suspension time, a corresponding sequence of pseudo labels can be calculated deterministically. The predicted RULs (pseudo labels) generated in this way satisfy the above constraints and the likelihood over the entire history is used to measure how likely the pseudo labels are correct. The other key idea is pseudo labeling utilizing the information obtained not only from the failure histories but also from the suspension histories. The information that only the failure histories have is poor, especially when the number of failure histories is limited. In this paper, the knowledge learned from the suspension histories is also used to improve the accuracy of the pseudo labels. Since suspension histories bring more diverse distribution in training data, it helps the model learn under more correct data distribution.

In the proposed method, two different neural networks (NN) with the same architecture are used. One is "NN for prediction of failure time" or "NN for pseudo labeling for suspension histories". We call it "teacher" because it gives answers to unlabeled data. The other is "NN that learns using both failure histories with true labels and suspension histories with pseudo labels assigned by teacher" which is called "student" because it studies with answers given by teacher. In the testing and practical use, we use only student, while teacher is used only in the training phase.

Figure 1 summarizes the overall learning process of the proposed method, which consists of two steps: (1) pre-training teacher only with failure histories, (2) training student with both failure and suspension histories. In step (1), there is only one sub-step (1a), the teacher is trained only with failure histories. In step (2), the student is trained using both failure and suspension histories, starting from the same parameters as the pre-trained teacher. Step (2) is further subdivided into three sub-steps: (2a) pseudo labeling, (2b) student training, and (2c) teacher update. We iterate these three sub-steps until the convergence. In sub-step (2a), pseudo labels are assigned to the suspension histories by the teacher. In sub-step (2b), the student is trained with failure histories and pseudo labeled suspension histories using the same training method as in (1a). Then, in sub-step (2c), the teacher's parameters are updated using the student's learning results in order to improve the teacher's performance. The details of each step are explained in the following subsections.

B. Construction of data

Before explaining the details of each step, we describe the assumed data format and its mathematical representation. Let $\mathcal{D}_{\mathrm{f}} = \{\mathcal{D}_{\mathrm{f},1}, ..., \mathcal{D}_{\mathrm{f},N}\}$ represent the N failure histories, where $\mathcal{D}_{\mathrm{f},i}$ is the i-th failure history, denoted by

$$\mathcal{D}_{\mathrm{f},i} = \{(\boldsymbol{x}_{\mathrm{f},i,1}, y_{\mathrm{f},i,1}), ..., (\boldsymbol{x}_{\mathrm{f},i,T_{\mathrm{f},i}}, y_{\mathrm{f},i,T_{\mathrm{f},i}})\}, \tag{1}$$

where $\boldsymbol{x}_{\mathrm{f},i,t} \in R^d$ and $y_{\mathrm{f},i,t}$ are d dimensional condition monitoring data and normalized RUL at time t respectively, and $T_{\mathrm{f},i}$ is the failure time of the i-th failure history. In addition, let $\mathcal{D}_{\mathrm{s}} = \{\mathcal{D}_{\mathrm{s},1}, ..., \mathcal{D}_{\mathrm{s},M}\}$ represent the M suspension histories, where $\mathcal{D}_{\mathrm{s},i}$ is the i-th suspension history, denoted by

$$\mathcal{D}_{\mathrm{s},i} = \{\boldsymbol{x}_{\mathrm{s},i,1}, ..., \boldsymbol{x}_{\mathrm{s},i,T_{\mathrm{s},i}}\}, \tag{2}$$

where $\boldsymbol{x}_{\mathrm{s},i,t} \in R^d$ is d dimensional condition monitoring data at time t, and $T_{\mathrm{s},i}$ is the suspension time of the i-th suspension history. Unlike the failure histories, RUL is not given. Instead, if the pseudo labels are assigned, these suspension histories can be utilized for the training in the same way as the failure histories.

As the models are supposed to take the temporal orders of the data samples into consideration in the (pre-)training phase, segments of a multidimensional time series are extracted as input by time windows of size r, described by

$$\boldsymbol{X}_{*,i,t} = (\boldsymbol{x}_{*,i,t-r+1}, ..., \boldsymbol{x}_{*,i,t}) \in R^{r\times d}, \tag{3}$$

where $*$ stands for f or s, and r is equal to the input size of the model. Then, both failure and suspension histories are rewritten by

$$\mathcal{D}_{\mathrm{f},i} = \{(\boldsymbol{X}_{\mathrm{f},i,1}, y_{\mathrm{f},i,1}), ..., (\boldsymbol{X}_{\mathrm{f},i,T_{\mathrm{f},i}}, y_{\mathrm{f},i,T_{\mathrm{f},i}})\}, \tag{4}$$

$$\mathcal{D}_{\mathrm{s},i} = \{\boldsymbol{X}_{\mathrm{s},i,1}, ..., \boldsymbol{X}_{\mathrm{s},i,T_{\mathrm{s},i}}\}. \tag{5}$$

C. Model architecture and training

Following the approach proposed in [3], the model used in this study outputs the predicted RUL as a distribution rather than a value. Specifically, it is assumed that the probability distribution of the RUL is normal distribution, and the model predicts its mean and variance, that is,

$$p(\hat{y}_{*,i,t}|\boldsymbol{X}_{*,i,t}) = \mathcal{N}(\mu(\boldsymbol{X}_{*,i,t}), \sigma^2(\boldsymbol{X}_{*,i,t})), \tag{6}$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ are the estimated mean and variance. This allows for pseudo labeling that takes into account the uncertainty of the prediction, and in practical use, it also allows for effective decision making that leverages the reliability of the prediction.

TABLE I
NEURAL NETWORK HYPERPARAMETERS.

Hyperparameter	Value
Receptive length	128
Encoder filters	11
Latent size	100
MLP hidden layers	1
MLP hidden size	100
Activation function	ReLU

While any architecture that satisfies this requirement, i.e., outputs mean and variance for pseudo labeling, can be used, in this study we used the same model as in [3], which is composed of two main components: an encoder and a RUL predictor. The input is embedded by the encoder of a 1D Convolutional Neural Network (CNN) with a large receptive field by the temporal dilation factor, and the RUL is then predicted from the embedded values by the RUL predictor of a simple Multilayer Perceptron (MLP). The hyperparameters of the model used in this paper are listed in Table I.

As in general deep learning, the model is trained by minimizing the loss function using optimization methods such as Stochastic Gradient Descent (SGD). Training using suspension histories can be done in the same way if pseudo labels can be properly assigned. In other words, the training in sub-step (1a) and sub-step (2b) is done in the same way. The loss function is assumed to be a negative log-likelihood with reference to [3]:

$$loss_* = \sum_{i,t}\left\{\frac{(\mu_{*,i,t} - y_{*,i,t})^2}{2\sigma^2_{*,i,t}} + \frac{1}{2}\ln\sigma^2_{*,i,t}\right\}. \tag{7}$$

However, since the loss of suspension histories is considered to be less reliable than that of failure histories due to the use of pseudo labels, by multiplying with a coefficient $\alpha < 1$ to reduce its relative importance, our semi-supervised loss can be described by

$$loss_{\mathrm{ss}} = loss_{\mathrm{f}} + \alpha \cdot loss_{\mathrm{s}}. \tag{8}$$

D. Pseudo labeling

One of the most important proposals of this study is how to assign pseudo labels to suspension histories, in the sub-step (2a). In this subsection, we explain the pseudo labeling for the

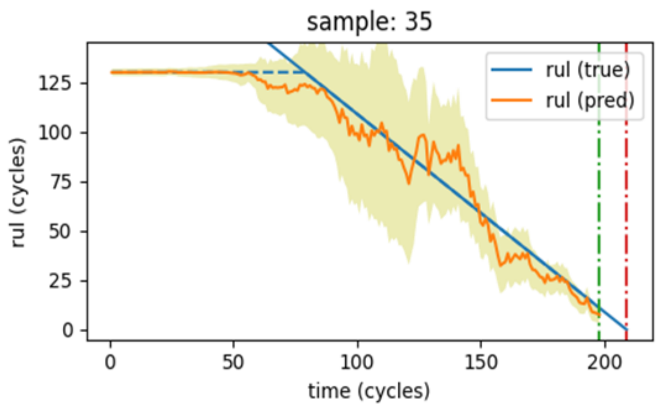

Fig. 2. An example of predicted RUL distribution sequence.

i-th suspension history, and the subscripts s and i are omitted when not needed.

The pseudo label at each time in a suspension history is the predicted RUL at each time, and we do not need to seek them independently if the failure time is available as we can calculate them all by counting backward from the failure time. For example, if the suspension time is assumed to be 10 and the predicted failure time is 12, the predicted RUL sequence is calculated as $\{11, 10, 9, 8, 7, 6, 5, 4, 3, 2\}$. This is one of the methods to calculate RULs from the failure time, and there are several developed methods to do this. In this study, a piece-wise RUL function proposed in [8] was used, which limits the maximum RUL to a constant value. In this case, for example, if the maximum RUL is assumed to be 8, the predicted RUL sequence is calculated as $\{8, 8, 8, 8, 7, 6, 5, 4, 3, 2\}$.

Thus, assuming K candidates of failure time which is greater than the suspension time, described by

$$\{\widetilde{T}_k | k = 1, ..., K; \widetilde{T}_k \geq T_{s,i}\}, \tag{9}$$

then we can obtain the corresponding candidates of the sequence of predicted RULs (pseudo labels), described by

$$\{\boldsymbol{L_k} | k = 1, ..., K\}, \tag{10}$$

where $\boldsymbol{L_k} = (L_{k,1}, ..., L_{k,T_{s,i}})$ and $L_{k,t}$ represents the pseudo label at time t of the k-th candidate. As mentioned in Section II-C, since the teacher model used in the proposed method outputs a predicted distribution of RUL at each time as Eq. (6), it is possible to calculate the log-likelihood for each pseudo label based on the distribution for each time. After the sequence of the distribution $(\mathcal{N}(\mu_1, \sigma_1^2), ..., \mathcal{N}(\mu_{T_{s,i}}, \sigma_{T_{s,i}}^2))$ is obtained as in Fig. 2 using Eq. (6), the log-likelihood for each pseudo label is calculated by

$$\mathcal{LL}_{k,t} = \log \mathcal{N}(L_{k,t} | \mu_t, \sigma_t^2). \tag{11}$$

Then, for a certain candidate of pseudo label sequence, the mean log-likelihoods of the pseudo labels over the entire history can be calculated as a score which represents how plausible this candidate is, expressed by

$$\overline{\mathcal{LL}_k} = \sum_{t=1}^{T_{s,i}} \mathcal{LL}_{k,t}. \tag{12}$$

After the scores of all K candidates are calculated, the optimal pseudo label sequence $\boldsymbol{L}_{\hat{k}}$ is selected by

$$\hat{k} = \arg\max_k \overline{\mathcal{LL}_k}, \tag{13}$$

and then the pseudo labeled suspension history is constructed as

$$\mathcal{D}_{s,i}^p = \{(\boldsymbol{x}_1, L_{\hat{k},1}), ..., (\boldsymbol{x}_{T_{s,i}}, L_{\hat{k},T_{s,i}})\}. \tag{14}$$

The pseudo labels assigned in this way not only satisfy two desirable conditions as RUL history described in Section II-A, but also can take into account the uncertainty of the prediction at each time by giving more weight to more confident points than to less confident points.

In the experiments in this study, actually, several reference time points were set as an equi-proportional sequence from the suspension time, and the mean log-likelihoods at the reference points were used as the score in order to reduce the computational cost.

E. Teacher update

The student model learns not only from the failure histories but also from the suspension histories with pseudo labels. On the other hand, the pseudo labels given by the teacher are not very reliable because the number of failure histories the teacher is pre-trained with is small. There is a risk of a negative effect on the performance improvement of the student who is trained with these pseudo labels. In order to improve the quality of the pseudo labels assigned by the teacher, we consider providing feedback to the teacher on the student's learning results, in sub-step (2c). A simple way to do this is to fully synchronize the student and the teacher, though there is no need to prepare two separate models in this case. Obviously, if the student is not learning well due to the poor pseudo labels, then providing feedback will lead to worse pseudo labeling. To prevent this, we consider teacher updating based on the loss of validation data. Since the performance evaluated on the validation data approximates the generalization performance, it is assumed that the student has good parameters when the loss of validation data decreases, and then the teacher is synchronized with the student at that time and not updated at other times. The following experiments and results discuss the difference in performance with these teacher update methods.

III. Experiments

A. Dataset description and preprocessing

The C-MAPSS dataset [9] provided by NASA is used to validate the RUL prediction performance of the proposed

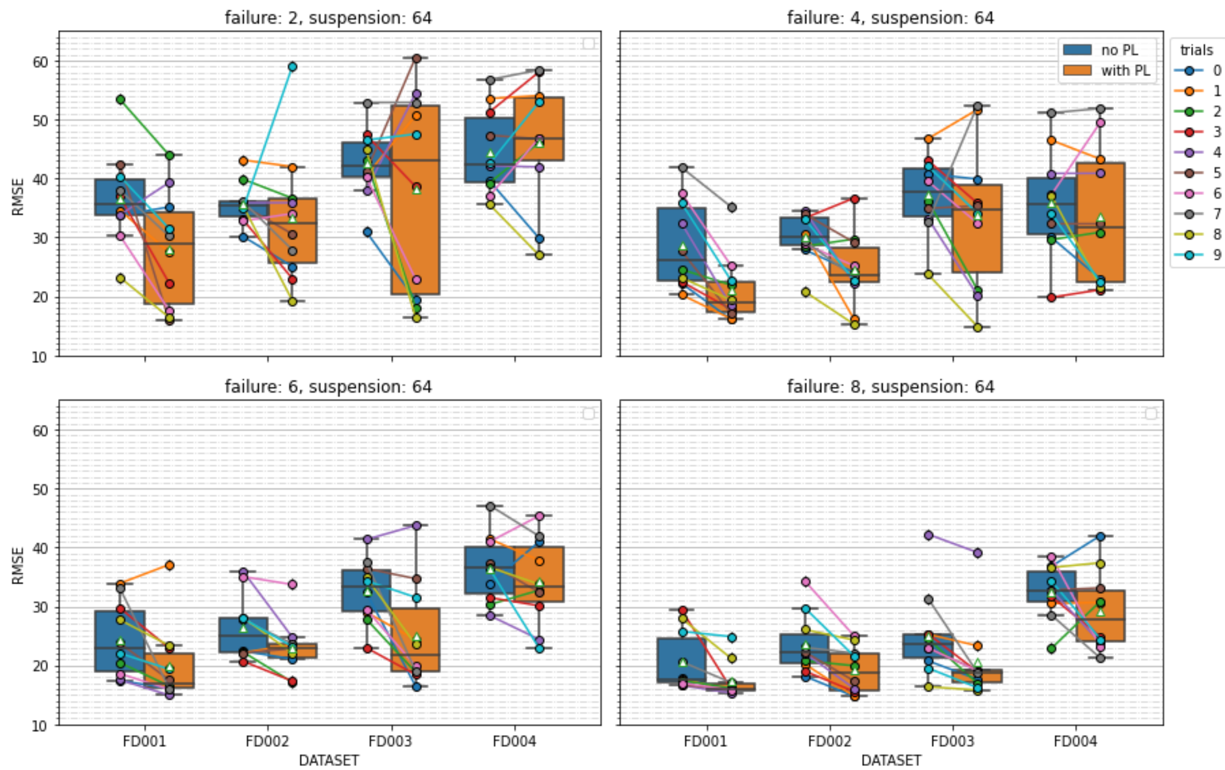

Fig. 3. RMSE boxplots of each 10 trials for each number of failure histories, for each subset, and without pseudo label learning (no PL) or with pseudo label learning (with PL).

TABLE II
THE C-MAPSS DATASET.

Subsets	FD001	FD002	FD003	FD004
Train engine number	100	260	100	249
Test engine number	100	259	100	248
Operation mode	1	6	1	6
Failure mode	1	1	2	2

method. It simulates actual turbofan engine degradation. The degradation history for each engine has 24 attributes, namely three operational settings and 21 sensors. This dataset contains four subsets FD001-FD004 with different operating conditions and failure modes, and each subset has training data and test data. In the training data, the engine runs to failure and the RUL of each cycle can be calculated. In the test data, on the other hand, the engine stops running before failure, but the RUL at the end of each history is provided. The details are shown in Table II.

In the preprocessing phase, we normalize the sensor measurements and the RUL values separately as follows. As to the sensors, measurements of each sensor are normalized with min-max normalization under the same operating conditions [10]. Since some of the sensor values do not change over time, only sensors 6, 7, 8, 11, 12, 15, and 17 are selected as input variables for the prediction. As to the RUL, following some studies using C-MAPSS dataset [6], [8], [10], [11], the maximum RUL is considered to be a constant value of 130, and the RUL greater than 130 is set to 130. Then the RUL is scaled within the range of $[0, 1]$ by dividing by 130. The suspension histories are generated from the training engines, randomly sampled from them, then the suspension time is randomly selected from 10-40 cycles before the failure time.

To evaluate the prediction accuracy, we use the root mean square error (RMSE), calculated as

$$RMSE = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}(RUL_{i,true} - RUL_{i,pred})^2}, \tag{15}$$

where $RUL_{i,true}$ and $RUL_{i,pred}$ are true and predicted RUL at the end of the i-th test history, and n_t is the number of test histories.

B. Results and discussions

First, in order to compare the performance in the case of training with failure histories only and in the case of training with pseudo labels as well, we experimented using all subsets from FD001 to FD004 with a different number of failure histories, set to 2, 4, 6, and 8, and 64 suspension histories. The coefficient α in the loss function Eq. (8) of sub-step (2b) is

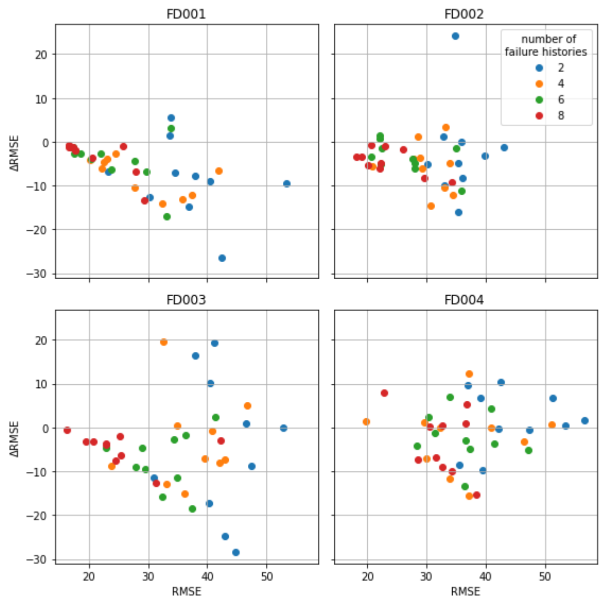

Fig. 4. Relationships between RMSE with only failure and ΔRMSE induced by pseudo label, and its difference in number of failure histories.

set to 0.8, and the teacher model is updated on every epoch in sub-step (2c). Since the maximum RUL is assumed to be 130, the set of possible candidates for the predicted failure time is $\{T_{s,i}+1, ..., T_{s,i}+130\}$. In sub-step (2a), K candidates are sampled from this set, and K is set to 120 at this time.

Figure 3 shows the boxplots of the RMSE obtained from 10 trials of each setting. In Fig. 3, the triangle points indicate the mean RMSE of each 10 trials, and the points connected by lines are the points where the same failure histories were used. The blue "no PL" means that the model was trained using only failure histories with no pseudo labels (no PL), and the orange "with PL" means that the model was trained using both failure histories and suspension histories with pseudo labels (with PL). Since the student learns using suspension histories with pseudo labels starting from the same parameters as the teacher which is pre-trained using only failure histories, the difference of RMSE between the corresponding "with PL" and "no PL" indicates the amount of improvement in RMSE due to pseudo label learning. In Fig. 3, it is observed that, comparing the mean RMSE (triangle points) between "no PL" and "with PL" for each subset with each number of failure histories, "with PL" records lower mean RMSE than "no PL" in every setting except FD004 with 2 failure histories. In other words, the performance improvement brought by pseudo label learning is confirmed at least statistically, although the degree of improvement varies in each trial. However, there are a small number of trials where the training using pseudo labels results in worse performance and larger RMSE. In Fig. 4, which is plotted with the RMSE at the end of the pre-training as the horizontal axis and the difference in RMSE due to pseudo label

TABLE III
NUMBER OF DETERIORATED TRIALS (IN 20 TRIALS).

	update by epoch	update by valid
FD001	2	2
FD002	4	1
FD003	8	7
FD004	11	8

learning, expressed as $\Delta RMSE$, as the vertical axis, it can be observed that the greater the subset number (from FD001 to FD004) or the smaller the number of failure histories, the more such cases of RMSE deterioration tend to occur. Specifically, in FD002 and FD003, when the number of failure histories is 2 and 4, there are trials where $\Delta RMSE$ is highly positive even though the RMSE at the end of pre-training is about the middle compared to other trials with the same number of failure histories, but not when the number of failure histories is 6 and 8. On the other hand, in FD004, perhaps because the task is too difficult due to the multiple operation modes and failure modes, there is no significant difference in the number of trials where the RMSE is deteriorated for each number of failure histories. Overall, however, it can be said that the larger the number of failure histories, the fewer the number of trials that deteriorate as a result of learning with pseudo labels, that is, the more consistently the performance is improved.

Next, in order to compare the two teacher update methods described in Section II-E, we trained the models using all subsets with 2 or 4 failure histories and 64 suspension histories. We call these two methods "update by epoch" and "update by valid" where the former is the way that the teacher is updated on every epoch and the latter is the way that the teacher is updated only when the loss on validation data records a minimum.

Figure 5 is the RMSE boxplots of 10 trials in each subset and each teacher update method, and Fig. 6 shows the RMSE improvement caused by pseudo label learning for each trial. These figures have the same format as Fig. 3 and Fig. 4, except that in Fig. 6, the black lines which mean the use of the same failure histories and the same pre-trained model are plotted. Table III shows the number of trials where the vertical axis value in Fig. 6 is greater than 0 for each subset and for each teacher update method. In other words, it indicates the number of trials where training using pseudo labels deteriorates the performance of the model. From Fig. 5, it can be seen that the teacher update method that performs better is different for each subset and each trial within that subset, and it is not clear which method is better even when comparing the mean RMSE of each 10 trials. However, Fig. 6 and Table III show that the number of trials where the performance is deteriorated due to the training using pseudo labels is less when "update by valid" is used than when "update by epoch" is used. Therefore,

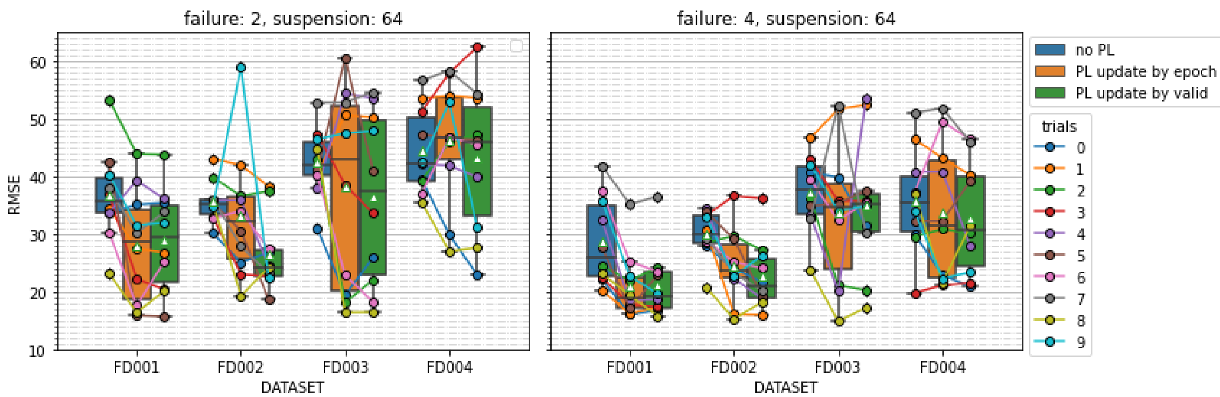

Fig. 5. RMSE boxplots of each 10 trials for each number of failure histories, for each subset, and for each teacher update method.

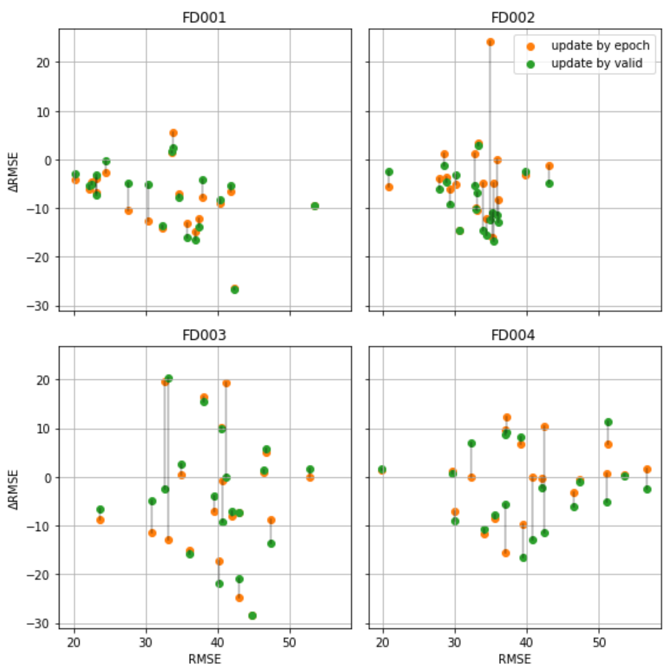

Fig. 6. Relationship between RMSE with only failure and ΔRMSE induced by pseudo label, and its difference in teacher update methods.

it can be said that "update by valid" is superior than "update by epoch" in the sense that it can improve model performance more stably.

IV. Conclusion

In this study, we proposed a semi-supervised RUL prediction method that utilizes both failure and suspension histories, a method that uses pseudo labels with a lower computational cost compared to previous studies using pseudo labels, while ensuring that constraints are satisfied. Experiments on the CMAPSS dataset showed that the use of pseudo labeled suspension histories allows more accurate RUL predictions than the use of failure histories alone, and provided several other insights, such as differences in results depending on the number of failure histories or the teacher update method.

One of the main contributions of this study is the pseudo labeling method which can assign pseudo labels to a suspension history based on likelihood, taking into account the entire history rather than just a part of it. In addition, this method allows the pseudo labels to naturally satisfy the constraints and to utilize the information from the suspension histories by teacher update. Furthermore, it has a high flexibility in terms of model architecture because it can be applied with any model whose input and output are the same as those of the model used in this paper. Although the method proposed in this paper is partly specific to the RUL prediction, it can be applicable to other tasks using time series data if developed further.

However, the experimental results show that under some conditions, the training with pseudo labels assigned by this method occasionally leads to worse performance of the model, especially in FD004. To solve this problem, we need to consider the various approaches, such as improving the teacher update method and developing the order of the data to be learned, like curriculum learning, and it is necessary to verify this as future work.

References

[1] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, Oct. 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0888327005001512.

[2] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems and Signal Processing*, vol. 107, pp. 241–265, Jul. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0888327017306064.

[3] J. Cornelius, B. Brockner, S. H. Hong, Y. Wang, K. Pant, and J. Ball, "Estimating and Leveraging Uncertainties in Deep Learning for Remaining Useful Life Prediction in Mechanical Systems," in *2020 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Jun. 2020, pp. 1–8.

[4] A. S. Yoon, T. Lee, Y. Lim, D. Jung, P. Kang, D. Kim, K. Park, and Y. Choi, "Semi-supervised Learning with Deep Generative Models for Asset Failure Prediction," Sep. 2017, arXiv:1709.00845 [cs]. [Online]. Available: http://arxiv.org/abs/1709.00845.

[5] C. Hu, B. D. Youn, T. Kim, and P. Wang, "A co-training-based approach for prediction of remaining useful life utilizing both failure and suspension data," *Mechanical Systems and Signal Processing*, vol. 62-63, pp. 75–90, Oct. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0888327015001004.

[6] R. He, Z. Tian, and M. J. Zuo, "A semi-supervised GAN method for RUL prediction using failure and suspension histories," *Mechanical Systems and Signal Processing*, vol. 168, p. 108657, Apr. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0888327021009833.

[7] Z. Tian, L. Wong, and N. Safaei, "A neural network approach for remaining useful life prediction utilizing both failure and suspension histories," *Mechanical Systems and Signal Processing*, vol. 24, no. 5, pp. 1542–1555, Jul. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088832700900377X.

[8] F. O. Heimes, "Recurrent neural networks for remaining useful life estimation," in *2008 International Conference on Prognostics and Health Management*, Oct. 2008, pp. 1–6.

[9] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *2008 International Conference on Prognostics and Health Management*, Oct. 2008, pp. 1–9.

[10] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long Short-Term Memory Network for Remaining Useful Life estimation," in *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Jun. 2017, pp. 88–95.

[11] M. Natsumeda and H. Chen, "RULENet: End-to-end Learning with the Dual-estimator for Remaining Useful Life Estimation," in *2020 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Jun. 2020, pp. 1–8.